Information Fusion 118 (2025) 102961

Contents lists available at ScienceDirect

Information Fusion

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

CD-UDepth: Complementary dual-source information fusion for underwater monocular depth estimation

Jiawei Guo [a], Jieming Ma [a,*], Feiyang Sun [a], Zhiqiang Gao [b], Ángel F. García-Fernández [c,d], Hai-Ning Liang [e], Xiaohui Zhu [a], Weiping Ding [f]

[a] School of Advanced Technology, Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, 215000, Jiangsu, China
[b] Department of Computer Science, College of Science, Mathematics and Technology, Wenzhou-Kean University, Wenzhou, 325060, Zhejiang, China
[c] Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, L69 3BX, UK
[d] ARIES center, Universidad Nebrija, Madrid, 28015, Spain
[e] Computational Media and Arts Thrust, Information Hub, the Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 511453, China
[f] School of Artificial Intelligence and Computer Science, Nantong University, Nantong, 226019, China

ARTICLE INFO

Dataset link: https://github.com/cainsmile/CD-UDepth

Keywords:
Monocular depth estimation
Underwater imaging
Information fusion
Deep learning

ABSTRACT

Underwater depth estimation is crucial for marine applications such as autonomous navigation and robotics. However, monocular depth estimation in underwater environments remains challenging due to the rapid attenuation of the red light spectrum in deep waters, causing bluish-green color distortion, while suspended particles and limited illumination lead to blurry effects. These underwater degradations severely affect the performance of RGB-based depth estimation methods, particularly in background regions. To overcome the limitations of color-based depth estimation techniques in underwater scenarios, this paper proposes a novel dual-source depth fusion framework leveraging color and light attenuation information. First, an innovative input space is designed inspired by the principle of depth-dependent light transmission in underwater environments. This input space enhances robustness against color distortion and improves the capacity to capture depth information, particularly in blurry underwater regions. Subsequently, we develop an adaptive fusion module to optimize the strengths of both RGB and this new input space across varying underwater conditions. This module employs a novel confidence-based mechanism to dynamically assess the reliability of depth information from each source on a per-pixel basis. By leveraging a learned confidence map, it can adaptively weigh and fuse the contributions of RGB and the new input space. This strategy enables optimal depth estimation across diverse underwater scenarios. Extensive experiments on multiple challenging datasets demonstrate that our method consistently outperforms current state-of-the-art monocular depth estimation techniques in various subaqueous environments.

1. Introduction

Underwater depth estimation is critical in ocean exploration applications, including subsea 3D reconstruction [1,2] and autonomous underwater vehicle (AUV) navigation [3,4]. Unlike terrestrial scenarios, underwater environments pose significant challenges for depth-sensing technologies such as LiDARs [5] and RGB-D cameras [6], involving high costs for deployment and maintenance, along with complex engineering requirements. Furthermore, the intrinsic complexities of underwater environments exacerbate sensing difficulties [7,8]. For example, the abundance of suspended particles and limited light conditions [9] can interfere with sensor signal reception, potentially leading to inaccuracies in depth measurements.

The emergence of deep learning has opened up new avenues for RGB image-based depth estimation research [10–16]. Normally, these approaches rely extensively on large-scale, high-quality terrestrial datasets such as NYU [17] and KITTI [18]. However, when applied to underwater imagery with significant chromatic distortions, these methods often struggle to achieve satisfactory outcomes. A significant challenge faced by these methods is the lack of real, high-quality underwater datasets. Many approaches rely on synthetic data [19] or transformed terrestrial images [20], which cannot fully capture the authentic characteristics of underwater environments. This limitation potentially restricts their ability to generalize to real-world underwater scenarios.

* Corresponding author.
E-mail address: Jieming.Ma@xjtlu.edu.cn (J. Ma).

https://doi.org/10.1016/j.inffus.2025.102961
Received 7 October 2024; Received in revised form 17 December 2024; Accepted 14 January 2025
Available online 21 January 2025

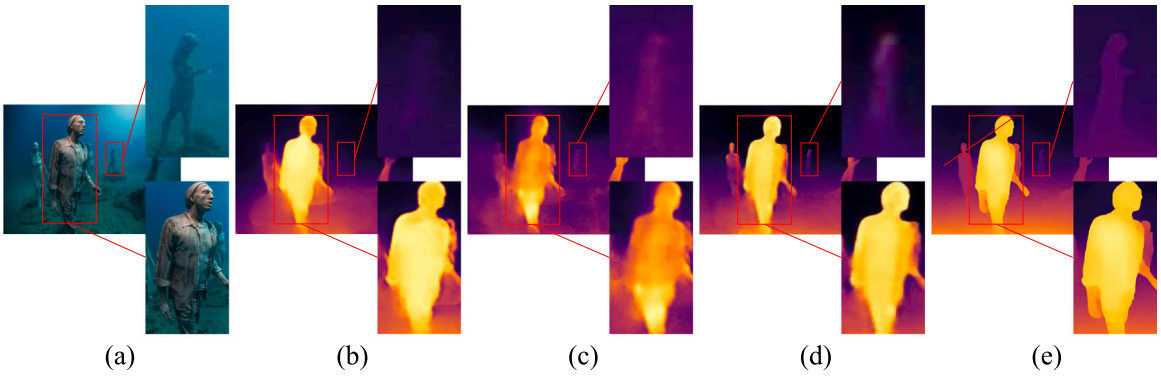

Fig. 1. Depth maps predicted by different sources: (a) RGB; (b) Depth map predicted by UDepth [21]; (c) Depth map predicted with IMT space as input; (d) Depth map predicted by the proposed CD-UDepth (ours). (e) Ground truth. Comparing with ground truth, depth from RGB provides higher accuracy in restoring the nearby, clearer statue (left sub-figure of each depth map), but almost completely fails to recover the distant, blurry statue (right sub-figure of each depth map). In contrast, depth from IMT clearly reveals the texture of the distant, blurry statue, but shows significant errors in restoring the nearby statue. Our CD-UDepth effectively combines these complementary strengths through adaptive fusion, producing depth maps that closely align with ground truth across both near and far regions.

Moreover, previous RGB-based methods occasionally fail to make accurate estimations for the object far from the camera and only perform well on closer regions with richer color and texture information, as shown in Fig. 1(b). This phenomenon can be attributed to the rapid attenuation of light as it travels through underwater environments, particularly within the red light spectrum, consequently leading to color distortion [22].

In this paper, we propose CD-UDepth, a novel complementary dual-source information fusion framework designed specifically for underwater scenes. To mitigate the vulnerability of RGB-based models to color distortion in underwater images, an input space IMT is designed. In this space, I represents grayscale intensity, M denotes the maximum value between the grayscale intensity and the green and blue channels, and T represents the underwater transmission. Transmission [23] is a coefficient for each pixel that reflects the variation in light intensity with underwater depth and is less susceptible to color distortion. Compared to RGB space, IMT space is more adept at extracting depth information from blurred background regions in underwater images. The IMT space exploits inherent characteristics of underwater imagery without dependence on external data.

To leverage the complementary strengths of both RGB and IMT spaces, we have developed two distinct depth estimation sub-modules: one tailored for RGB input and another for IMT input. These modules are designed to extract depth information from their respective input spaces. However, a simple combination of these outputs fails to fully capitalize on the unique advantages offered by each space. RGB space typically excels in areas with color-rich foreground objects, as shown in Fig. 1(b). IMT space proves more effective in regions with blurred textures, often found in the background of underwater scenes, as shown in Fig. 1(c). To effectively integrate these complementary features, we propose a confidence matrix learning strategy that adaptively evaluates the reliability of depth cues from different sources. This strategy is implemented through our complementary dual-source information fusion module (CDIFM). The confidence matrix assesses the reliability of depth information from each source and dynamically adjusts the fusion weights accordingly, giving preference to the more reliable source in each region of the image. The effectiveness of this approach is evident when comparing the outputs. As illustrated in Fig. 1(d), the proposed CD-UDepth method successfully combines the strengths of both spaces, resulting in depth maps that demonstrate improved accuracy and completeness across varied underwater scenes.

The contributions of this paper can be summarized as follows:

- To mitigate the loss of depth information due to color distortion in underwater environments, a novel IMT input space guided by light attenuation priors is introduced in this paper. This space combines grayscale intensity, maximum value between grayscale and blue–green channels, and transmission rate, which provides a more reliable depth indicator by capturing the relationship between light attenuation and depth.
- To synergize the contributions of color and light attenuation information, we design a complementary dual-source information fusion module. This module incorporates a confidence matrix learning strategy that dynamically evaluates the reliability of depth estimates from both RGB and IMT spaces. By learning confidence scores that reflect the trustworthiness of each source's depth cues, our module adaptively fuses information from different sources to achieve optimal depth estimation under varying underwater conditions.
- Building upon the above contributions, this paper proposes the CD-UDepth framework for underwater depth estimation, realizing the joint modeling and utilization of complementary information sources. The framework effectively combines the advantages of both spaces: the strength of RGB in processing well-lit regions and the robustness IMT in handling degraded areas. Extensive experiments demonstrate that CD-UDepth achieves state-of-the-art performance across various underwater conditions.

The remainder of this paper is organized as follows: Section 2 introduces related work on monocular depth estimation and underwater imagery. In Section 3, we provide a detailed description of CD-UDepth. Section 4 presents experimental results comparing CD-UDepth with various state-of-the-art algorithms in underwater scenes, ablation studies, and robustness analysis. Finally, conclusions are drawn in Section 5.

2. Related work

2.1. Deep learning-based monocular depth estimation

Deep learning has substantially improved monocular depth estimation by utilizing encoder–decoder architectures to learn image representations and generate the corresponding depth maps. This approach was initially proposed by Eigen et al. [24] in 2014. With the advancement of deep learning, increasing depth estimation models have been proposed to address a variety of challenges. Song et al. [25] propose a simple but effective network to predict depth by incorporating the Laplacian pyramid into the decoder architecture. Adabins [14] presents a module that utilizes the Transformer-based approach [26] to partition the depth range into multiple bins and estimate the final depth values via linear combinations of bin centers. Since then, many researchers have adopted Adabins and Transformer as the backbone structure for depth estimation models [10–13,27]. Faced with the scarcity of paired datasets, semi-supervised and unsupervised techniques have

been introduced. These approaches deduce scene depth from geometric constraints in multi-view imagery and handle sequences of images [28] and videos [29,30]. Moreover, the integration of photometric [31] and symmetry losses [32] has been proposed to improve model-constrained depth accuracy. Xiang et al. [33] present a self-supervised multi-frame depth estimation framework that enhances camera pose estimation by fusing visual and inertial modalities using a novel visual-inertial fusion Transformer, and introduces monocular depth priors to adaptively modulate the multi-frame cost volume features

There is relatively little research on the robustness of depth estimation in the presence of image degradation and complex environments. Addressing the challenge of image distortion in panoramic depth estimation, Chen et al. [34] propose the distortion-aware monocular omnidirectional (DAMO) network, which effectively extracts semantic features from distorted panoramas and leverages a spherical-aware weight matrix to handle uneven area distribution. Hanjiang et al. [35] introduce the SeasonDepth dataset and benchmark to evaluate the robustness of learning-based monocular depth estimation methods across diverse environmental conditions, providing insights for improving autonomous driving in complex real-world scenarios. Lingdong et al. [36] introduce the RoboDepth benchmark to comprehensively assess the robustness of learning-based monocular depth estimation models against a wide range of common corruptions, underscoring the need for dedicated robustness evaluation and informing the design considerations for crafting more resilient depth perception systems. Long et al. [37] propose a model for robust depth completion from sparse and non-uniform inputs, which leverages a stable feature fusion module and an uncertainty-based feature embedder to effectively handle poor quality depth maps in real-world usage. While these methodologies are trailblazing and perform remarkably well on high-quality terrestrial datasets, their robustness fails in the challenging conditions of underwater environments.

2.2. Underwater imagery and application

Research on underwater imagery has primarily focused on developing techniques for underwater image enhancement and restoration [38–42] to address the unique factors of the underwater environment, as well as methods for underwater object detection and recognition [43–45]. For instance, Fayaz [43] propose SwinWave-SR, a novel multi-scale vision Transformer-based algorithm that leverages wavelets to enable efficient and accurate super-resolution of underwater images by preserving key high-frequency components and reducing computational cost. In addition, some researchers have also been working on the fields of underwater autonomous driving [4] and underwater robotics [46]. Pinto et al. [47] design an innovative hybrid underwater imaging system that combines active and passive techniques to provide dense and accurate 3D information less affected by harsh underwater conditions compared to conventional methods, enabling reliable 3D data for maritime inspection and autonomous underwater navigation. In these works, underwater depth perception plays an important role.

2.3. Underwater depth estimation

Current leading methods for underwater depth detection predominantly involve active approaches incorporating sensor integration [9,48,49]. Image-based underwater depth estimation typically incorporates the dark channel prior [50] and transmission as prior knowledge [20], such as dark channel prior (DCP) [51] and underwater dark channel prior (UDCP) [23]. Considering the underwater environment, the red wavelength experiences the fastest attenuation. Adrian et al. [52] utilize the red channel as a variant of the dark channel to estimate underwater depth. Additional information, such as blurriness [53] and spectral profiles [54], are also considered as factors in describing underwater depth. To simulate underwater images, Gupta et al. [19] establish a neural network model to learn the mappings between hazy underwater and above-water hazy color images and depth maps. Hambarde et al. [20] synthesize underwater images based on adversarial generative networks, using ground images. Yu et al. [21] develop a lightweight underwater depth estimation model, UDepth, based on Adabins [14], using a least-squared formulation for coarse pixel-wise depth prediction. However, their method has a high dependency on datasets and lacks generalization capability across different datasets. Chen et al. [55] propose a physical-guided Transformer-based underwater monocular depth estimation method that integrates underwater imaging characteristics and physical priors through multiple specialized modules to achieve superior depth estimation performance in degraded underwater environments. However, existing methods often struggle to accurately estimate the depth of distant objects, as light attenuation and scattering effects become more pronounced with increasing distance, resulting in severe loss of feature information for far-away objects. These limitations highlight the necessity of developing more robust and adaptive underwater depth estimation methods.

3. Method

To address the challenges of accurate depth estimation in underwater environments, where color distortion and light attenuation significantly impact traditional methods, a novel depth estimation framework CD-UDepth is proposed, as illustrated in Fig. 2. In this chapter, Section 3.1 introduces the proposed IMT input space, and Section 3.2 delineates the depth estimation sub-model employed in the method. Section 3.3 provides a detailed exposition of the complementary dual-source information fusion module. Finally, Section 3.4 elucidates the loss functions utilized during the training process.

3.1. IMT space

In underwater images, the red color channel is subject to intense attenuation, leading to its concentration within a narrow, low-end range of the spectrum [56,57]. Although the red channel contains valuable information in shallow water regions where red light has not been fully attenuated, its reliability significantly decreases with increasing depth due to wavelength-dependent attenuation characteristics in water. The IMT space is specifically designed to address this unique characteristic of underwater image degradation by combining three complementary components: grayscale intensity, maximum value between green and blue channels, and transmission.

Grayscale intensity preserves structural information through luminance, which has been proven effective in underwater feature detection [58]. The maximum value between green and blue channels is utilized based on the observation that these wavelengths exhibit different penetration characteristics in water — green light penetrates better in coastal waters while blue light shows advantages in oceanic waters [59]. This selective combination helps maintain the strongest available signals for depth estimation.

The third component, transmission, reflects the degree to which light intensity diminishes with increasing underwater depth [57]. The principle known as DCP posits that in most regions of non-sky optical images, at least one color channel contains pixels with near-zero intensity levels [51]. UDCP [23] extends the dark channel prior to underwater scenes by considering the specific light attenuation characteristics in underwater environments. Due to the wavelength-dependent attenuation of light in water, the image formation process can be described using a widely-adopted model:

$$I(x) = J(x)T(x) + A(1 - T(x)), \tag{1}$$

where I is the observed image, J is the scene radiance, A is the ambient light, and T is the transmission.

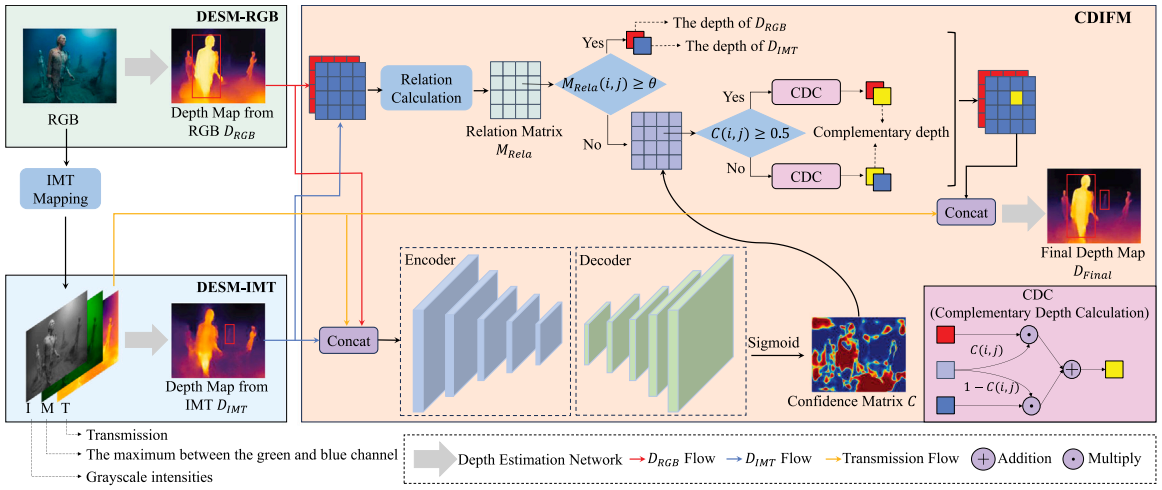

Fig. 2. Overview of the proposed CD-UDepth framework. It first predicts depth maps from color and light attenuation-guided information by two depth estimation sub-modules, DESM-RGB and DESM-IMT, respectively. A confidence-guided fusion module CDIFM then adaptively combines the outputs from both sources, producing the final depth prediction.

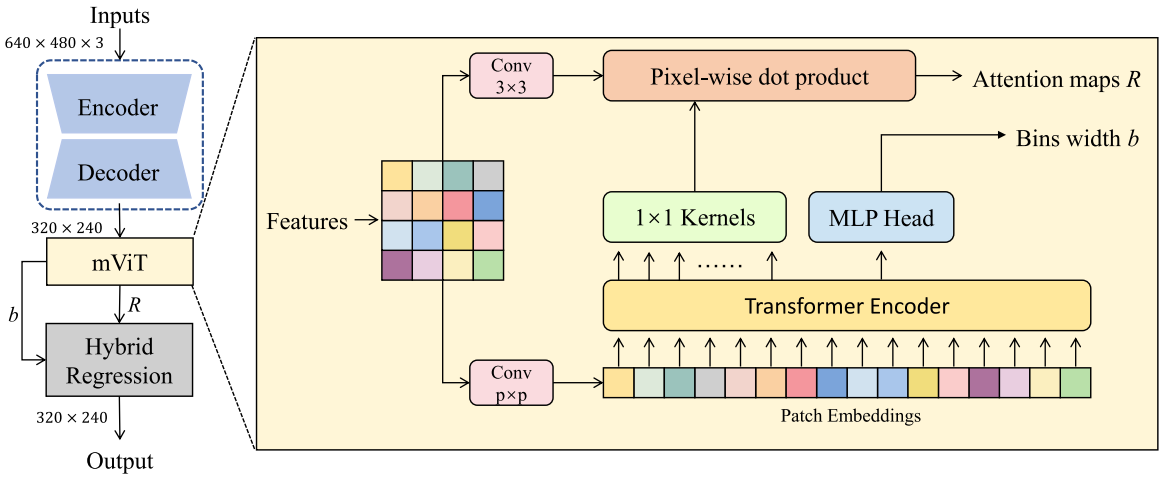

Fig. 3. The architecture of the depth estimation Sub-module.

Unlike the standard DCP, UDCP specifically addresses underwater scenarios by excluding the red channel from consideration, as red light attenuates much more rapidly in water compared to green and blue light. Based on this principle, the transmission can be estimated as:

$$\tilde{T}(x) = 1 - \min_{y \in \Omega(x)} \left(\min_{c \in G,B} \frac{I^c(y)}{A^c} \right), \tag{2}$$

where $\Omega(x)$ is a local patch centered at x, and G, B represent the green and blue channels respectively.

3.2. Depth estimation sub-module

In CD-UDepth, depth estimation sub-modules (DESM) are constructed to learn the mappings from the input spaces to depth maps, and are named DESM-RGB and DESM-IMT respectively, based on their corresponding input spaces. The depth estimation methodology follows the designs of Adabins [14] and UDepth [21] by incorporating two pivotal components: a streamlined encoder–decoder backbone, an efficient vision Transformer, and a convolutional regression module. Fig. 3 provides a comprehensive overview of the architecture.

The initial component of the proposed network employs a computationally efficient encoder–decoder backbone. We have opted for a lightweight model based on MobileNetV2 [60] for the encoder, which offers improved efficiency compared to more complex alternatives. It utilizes an inverted residual structure and starts with 32 filters, culminating in 48 filters through 19 bottleneck layers, to produce depth maps at 320 × 240 resolution. This choice is motivated by the need for real-time performance in underwater applications. The encoder–decoder module adopts a U-Net-like structure. The encoder processes the input RGB image through multiple stages, extracting features at various levels for use in skip connections. The decoder comprises several upsampling layers that integrate the current features with encoder skip connections and appropriately scaled depth prior parameterizations.

To complement the spatial features extracted by the encoder–decoder, a miniaturized visual Transformer (mViT) [26] is incorporated to capture global context. The approach incorporates a multi-layer perceptron (MLP) that processes the initial embedding output. This MLP utilizes a rectified linear unit activation function and generates a vector of N dimensions, referred to as the preliminary bin-width vector. The next step in the process involves normalizing the preliminary vector. The values are adjusted so that their sum equals one, effectively creating a probability distribution. The formulation for computing the bin widths b_i is as follows:

$$b_i = \frac{b'_i + \epsilon}{\sum_{j=1}^{N} \left(b'_j + \epsilon \right)}, \tag{3}$$

where b'_i is predicted through the MLP head from the first Transformer output embedding, and the small positive $\epsilon = 10^{-3}$ ensures each bin width is strictly positive.

The hybrid regression module, which forms the final stage of the network, constructs the depth image prediction by leveraging bin centers and range attention maps generated by preceding modules. This process involves several computational steps to refine raw outputs into a coherent depth estimate. Initially, the system calculates bin classification scores $p_i(x, y)$ for each pixel (x, y). This computation applies a compact 1×1 convolution to the range attention maps, followed by a softmax activation. The result is a set of n probabilities for each pixel, where n represents the total number of bins. The final depth prediction $\hat{d}(x, y)$ for each pixel is then formulated as a weighted sum:

$$\hat{d}(x, y) = \sum_{i=1}^{n} p_i(x, y) \cdot c_i, \tag{4}$$

where $p_i(x, y)$ represents the probability of pixel (x, y) belonging to bin i, and c_i denotes the center depth value of bin i. This approach effectively combines discrete bin probabilities with their corresponding depth values to produce a continuous depth estimate.

3.3. Complementary dual-source information fusion module

To leverage the advantages of both color texture and light attenuation, this paper proposes a complementary dual-source information fusion module (CDIFM), as illustrated in Fig. 2. The strategy of CDIFM is presented in detail in Section 3.3.1. Subsequently, Section 3.3.2 delineates the learning methodology for the confidence matrix, which is employed to weight the two input spaces within the framework.

3.3.1. Strategy of CDIFM

Assuming that depth maps $\mathbf{D}_{\mathrm{RGB}}$ and $\mathbf{D}_{\mathrm{IMT}}$ have been obtained from RGB and IMT spaces respectively, we first calculate the relationship matrix $\mathbf{M}_{\mathrm{Rela}}$ by taking the absolute difference between $\mathbf{D}_{\mathrm{RGB}}$ and $\mathbf{D}_{\mathrm{IMT}}$:

$$\mathbf{M}_{\mathrm{Rela}}(x, y) = 1 - |\mathbf{D}_{\mathrm{RGB}}(x, y) - \mathbf{D}_{\mathrm{IMT}}(x, y)|, \tag{5}$$

where (x, y) denotes the pixel coordinate. To ensure unbiased and consistent integration with subsequent fusion steps, we apply min–max normalization to the relationship matrix:

$$\mathbf{M}_{\mathrm{Rela}}(x, y) = \frac{\mathbf{M}_{\mathrm{Rela}}(x, y) - \min(\mathbf{M}_{\mathrm{Rela}})}{\max(\mathbf{M}_{\mathrm{Rela}}) - \min(\mathbf{M}_{\mathrm{Rela}})}. \tag{6}$$

Given a predefined threshold θ, when $\mathbf{M}_{\mathrm{Rela}}(x, y) \geq \theta$, the depth values at (x, y) from both $\mathbf{D}_{\mathrm{RGB}}$ and $\mathbf{D}_{\mathrm{IMT}}$ are directly concatenated. This high agreement between the two depth estimations suggests that both depth values at that pixel are reliable. When $\mathbf{M}_{\mathrm{Rela}}(x, y) < \theta$, indicating significant disagreement between RGB and IMT depths, the fusion process is governed by the confidence matrix $\mathbf{C}$, which produces values between 0 and 1. Values closer to 1 indicate higher reliability in RGB depth estimates, while values closer to 0 indicate higher reliability in IMT depth estimates. The threshold 0.5 serves as the natural midpoint to determine the arrangement of fusion channels: when $\mathbf{C}(x, y) \geq 0.5$, RGB depth is set as the primary source, while when $\mathbf{C}(x, y) < 0.5$, IMT depth becomes the primary source. Rather than discarding the less confident source, a complementary depth is computed through weighted combination:

$$d_{\mathrm{Com}} = \mathbf{C}(x, y) \cdot \mathbf{D}_{\mathrm{RGB}}(x, y) + (1 - \mathbf{C}(x, y)) \cdot \mathbf{D}_{\mathrm{IMT}}(x, y), \tag{7}$$

where d_{Com} is the complementary depth at (x, y). This weighted combination preserves potential useful information from both sources while maintaining the priority of the more confident source.

Finally, as most features in the dual-source information are color features, the representation of light attenuation effects is enhanced by concatenating the transmission with the complementary dual-source depth map. This serves as the input for the final depth estimation model, which is identical to that introduced in Section 3.2. The steps for implementing the CDIFM strategy are provided in Algorithm 1.

Algorithm 1 Strategy of CDIFM

Require: Depth maps $\mathbf{D}_{\mathrm{RGB}}$, $\mathbf{D}_{\mathrm{IMT}}$, confidence matrix $\mathbf{C}$, threshold θ, transmission $\mathbf{T}$
Ensure: Final depth map $\mathbf{D}_{Final}$

```
for all pixels (i, j) do
    M_Rela(x, y) = |D_RGB(x, y) − D_IMT(x, y)|
end for
M_Rela(x, y) = (M_Rela(x, y) − min(M_Rela)) / (max(M_Rela) − min(M_Rela))
for all pixels (i, j) do
    if M_Rela(x, y) < θ then
        d_Com = C(x, y) · D_RGB(x, y) + (1 − C(x, y)) · D_IMT(x, y)
        if C(x, y) ≥ 0.5 then
            F_Com(x, y) = [D_RGB(x, y), d_Com]
        else
            F_Com(x, y) = [d_Com, D_IMT(x, y)]
        end if
    else
        F_Com(x, y) = [D_RGB(x, y), D_IMT(x, y)]
    end if
end for
D_Final = DepthEstimation([F_Com(x, y), T(x, y)])
return D_Final
```

3.3.2. Confidence matrix learning

In the CDIMF, we employ an encoder–decoder network to learn the confidence matrix. The encoder extracts multi-scale features from the input image, while the decoder gradually reconstructs a high-resolution feature map. The decoder utilizes a series of upsampling blocks, each containing 3×3 convolutions and LeakyReLU activations, before upscaling the spatial resolution to match corresponding encoder outputs. Skip connections between encoder and decoder layers enhance feature integration, culminating in a one-channel output through a sigmoid function to represent confidence scores for dual-source depth maps. The network input is a three-channel feature map concatenated as $[\mathbf{D}_{\mathrm{RGB}}, \mathbf{D}_{\mathrm{IMT}}, \mathbf{T}]$, where $\mathbf{T}$ is the transmission for all pixels. To generate a pseudo-label $\mathbf{P}$, we calculate the error between $\mathbf{D}_{\mathrm{RGB}}$ and $\mathbf{D}_{\mathrm{IMT}}$, compared with the ground truth $\mathbf{D}_{GT}$, as follows:

$$\mathbf{P}(i, j) = \begin{cases} 1 & \text{if } |\mathbf{D}_{\mathrm{RGB}}(x, y) - \mathbf{D}_{\mathrm{GT}}(x, y)| \geq |\mathbf{D}_{\mathrm{IMT}}(x, y) - \mathbf{D}_{\mathrm{GT}}(x, y)|, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The learned confidence values range between 0 and 1, where a value closer to 1 indicates higher confidence in the depth map $\mathbf{D}_{\mathrm{RGB}}$, while a value closer to 0 indicates higher confidence in $\mathbf{D}_{\mathrm{IMT}}$.

3.4. Loss function

The training loss incorporates four components: L_{RGB} and L_{IMT} for the mapping from RGB space and IMT space to the depth map, L_{F} for the final depth estimation model, and $L_{\mathbf{C}}$ for the confidence matrix learning model. Among these, L_{RGB}, L_{IMT}, and L_{Final} serve as loss functions for the depth estimation model, and they share the same model structure. Hence, they employ the same loss functions, named L_{Depth}.

3.4.1. Loss function for depth estimation

In this paper, the loss function for depth estimation L_{Depth} comprises two fundamental components:

1. **Scale-Invariant Logarithmic (SILog) loss.** For underwater tasks, a variant of the SILog loss [61] is initiated to strike a balance between metric precision and relative depth estimation. Let $\mathbf{D}$ and $\hat{\mathbf{D}}$ be the true and predicted depth maps, respectively.

Then the SILog loss is adjusted as:

$$L_{\text{SILog}}(\mathbf{D},\hat{\mathbf{D}}) = \sqrt{\frac{1}{N}\sum_{(x,y)} g^2(x,y) - \frac{\alpha}{N^2}\left(\sum_{(x,y)} g(x,y)\right)^2}. \tag{9}$$

Here, $g(x,y)$ denotes the logarithmic space error, defined as $g(x,y) = \log \hat{\mathbf{D}}(x,y) - \log \mathbf{D}(x,y)$, in which $\mathbf{D}(x,y)$, and $\hat{\mathbf{D}}(x,y)$ represent the value of $\mathbf{D}$ and $\hat{\mathbf{D}}$ at (x,y), respectively. N is the number of pixels, and α is a balancing factor, set to 0.15 in our experiments.

2. **Mean Squared Error (MSE) loss.** We employ the standard MSE loss to encourage the model to make predictions that match the metric scale of the true depths. The MSE loss is formulated as:

$$L_{\text{MSE}}(\mathbf{D},\hat{\mathbf{D}}) = \frac{\beta}{N}\sum_{i}(\hat{\mathbf{D}}(x,y) - \mathbf{D}(x,y))^2, \tag{10}$$

where β is set to 10 to balance the weight of the components.

3. **Chamfer Distance Loss.** This component of the loss function aims to align the distribution of bin centers with the distribution of actual depth values in the ground truth data. The objective is to promote a close correspondence between the bin centers and the true depth values, operating in both directions. Following adabins, the chamfer distance loss [62] is calculated as:

$$L_{\text{Chamfer}}(c,\mathbf{D}) = \sum_{c_i\in c}\min_{\mathbf{D}(x,y)\in\mathbf{D}} \left\|\mathbf{D}(x,y) - c_i\right\|_2^2 + \sum_{\mathbf{D}(x,y)\in\mathbf{D}}\min_{c_i\in c}\left\|\mathbf{D}(x,y) - c_i\right\|_2^2, \tag{11}$$

in which c is the set of bin centers.

The final loss for depth estimation L_{Depth} can be expressed as the sum of the two components:

$$L_{\text{Depth}}(c,\mathbf{D},\hat{\mathbf{D}}) = L_{\text{SILog}}(\mathbf{D},\hat{\mathbf{D}}) + L_{\text{MSE}}(\mathbf{D},\hat{\mathbf{D}}) + L_{\text{Chamfer}}(c,\mathbf{D}). \tag{12}$$

3.4.2. Loss function for confidence matrix learning

In the training of the confidence matrix learning, where the pseudo label $\mathbf{P}$ is a matrix with elements of 0 or 1, the loss function is the combination of MSE loss and binary cross-entropy (BCE) loss which are suitable for binary classification. Given the estimated confidence matrix $\mathbf{C}$, the BCE loss can be calculated as:

$$L_{\text{BCE}}(\mathbf{C},\mathbf{P}) = -\frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{P}(x,y)\log(\mathbf{C}(x,y)) + (1-\mathbf{P}(x,y))\log(1-\mathbf{C}(x,y))\right], \tag{13}$$

in which $\mathbf{C}(x,y)$, and $\mathbf{P}(x,y)$ represent the value of $\mathbf{C}$ and $\mathbf{P}$ at (x,y), respectively. Therefore, the loss function for confidence matrix learning $L_{\mathbf{C}}$ can be expressed as:

$$L_{\mathbf{C}} = L_{\text{BCE}}(\mathbf{C},\mathbf{P}) + L_{\text{MSE}}(\mathbf{C},\mathbf{P}). \tag{14}$$

4. Experiments and results

4.1. Implementation settings

The model was implemented with PyTorch 1.7.1, leveraging an NVIDIA RTX A5000 GPU with CUDA 11.0 for computational acceleration. The training process follows a three-stage strategy. In the initial stage, the RGB-based and IMT-based depth estimation sub-networks are independently trained using the Adam optimizer with a learning rate of 1.5×10^{-4}. Subsequently, we freeze these pre-trained depth estimation sub-networks and train the confidence prediction autoencoder to learn optimal fusion weights. In the final stage, the entire network, including the final depth estimation sub-network, is fine-tuned end-to-end with the same optimizer settings. Each stage was trained for 50 epochs with a batch size of 4, with model selection based on the best validation set performance.

4.2. Dataset

To evaluate our proposed method, we conducted experiments on two comprehensive datasets. These datasets are introduced as follows:

USOD10K [63] contains paired images featuring 70 categories of salient objects across 12 underwater scenes. It includes RGB images and corresponding ground truth depth maps at a resolution of 640 × 480 pixels, with 7178 training and 1026 testing samples.

FLSea [64] dataset, collected and developed by the University of Haifa in Israel, comprises paired images and depth maps from multiple scenes across two regions: Canyons in the Mediterranean Sea and the Red Sea. The Canyons dataset encompasses four scenes: U Canyon, Horse Canyon, Tiny Canyon, and Flatiron, consisting of 2875, 2475, 1012, and 2230 frames, respectively. The Red Sea dataset includes eight scenes: Big Dice Loop, Coral Table Loop, Cross Pyramid Loop, Dice Path, Northeast Path, Landward Path, Pier Path, and Sub Pier. In this study, we randomly partition The Canyons dataset, allocating 70% for training and 30% for testing to evaluate the performance of CD-UDepth. Furthermore, to assess cross-regional robustness, we conduct tests on three scenes from the Red Sea region, specifically Big Dice Loop, Coral Table Loop, and Sub Pier.

4.3. Evaluation metrics

This paper evaluates the CD-UDepth using quantitative metrics that are standard for assessing depth estimation models [21], including absolute relative error (Abs Rel), root mean squared error (RMSE), $\log_{10}$ error ($\log_{10}$), and squared relative error (Sq Rel). These metrics indicate the error between the predicted depth map and the true depth map; the smaller they are, the higher the quality of the predicted depth map.

- Threshold accuracy (δ_i): $max\left(\frac{d_p}{\hat{d}_p}, \frac{\hat{d}_p}{d_p}\right) = \delta < \Gamma$ for $\Gamma = 1.25, 1.25^2, 1.25^3$;
- Abs Rel: $\frac{1}{n}\sum_{p=1}^{n}\frac{\left|d_p-\hat{d}_p\right|}{d}$;
- RMSE: $\sqrt{\frac{1}{n}\sum_{p=1}^{n}\left(d_p-\hat{d}_p\right)^2}$;
- $\log_{10}$ error: $\frac{1}{n}\sum_{p=1}^{n}\left|\log_{10}(d_p)-\log_{10}(\hat{d}_p)\right|$;
- Sq Rel: $\frac{1}{n}\sum_{d}^{n}\frac{\left\|d_p-\hat{d}_p\right\|^2}{d}$,

where d_p is a pixel in depth image d, $\hat{d}_p$ is a pixel in the predicted depth image $\hat{d}$, and n is the total number of pixels in d. Γ is the threshold value that defines the accuracy criterion.

4.4. Comparison to state-of-the-art algorithms

To the best of our knowledge, UDepth [21] is the first method to perform depth estimation on the large-scale paired underwater dataset. UPGformer [55], UW-LapDepth [25], and UDepth are all specifically designed for underwater depth estimation using paired training data. Due to the unavailability of the source code of UPGformer, we cite the results from their paper but do not include their visualized depth estimation results. In order to contrast the advantages of CD-UDepth in underwater environments, we also reproduce some advanced depth estimation models on the underwater datasets, including IEbins [10], NeWCRFs [15], PixelFormer [11], URCDC-Depth [65], and VA-DepthNet [66].

Table 1 presents a comparison between the proposed CD-UDepth and state-of-the-art methods on the USOD10K dataset [63]. CD-UDepth demonstrates superior performance across most evaluation metrics. For threshold accuracy metrics, our method achieves 0.496, 0.720, and 0.833 respectively, showing significant improvements of 12.5%, 5.1%, and 2.3% over the previous best method UW-LapDepth. Although UPGformer achieves marginally better performance in Abs Rel,

Table 1
Qualitative comparison to SOTA methods on the USOD10K dataset [63].

Method	δ_1 ↑	δ_2 ↑	δ_3 ↑	Abs Rel ↓	RMSE ↓	$\log_{10}$ ↓	Sq Rel ↓
IEbins [10]	0.139	0.292	0.478	1.211	0.288	0.326	0.262
NeWCRFs [15]	0.249	0.462	0.623	0.814	0.237	0.281	0.179
PixelFormer [11]	0.377	0.611	0.752	0.981	0.180	0.214	0.198
URCDC-Depth [65]	0.229	0.426	0.580	0.797	0.253	0.306	0.182
VA-DepthNet [66]	0.245	0.504	0.699	1.434	0.215	0.262	0.327
UPGformer [55]	0.442	0.677	0.809	**0.525**	0.146	0.177	0.100
UW-LapDepth [25]	0.441	0.685	0.814	0.681	0.151	0.179	0.150
UDepth [21]	0.352	0.612	0.772	0.681	0.143	0.202	0.195
CD-Udepth	**0.496**	**0.720**	**0.833**	0.536	**0.127**	**0.165**	**0.086**

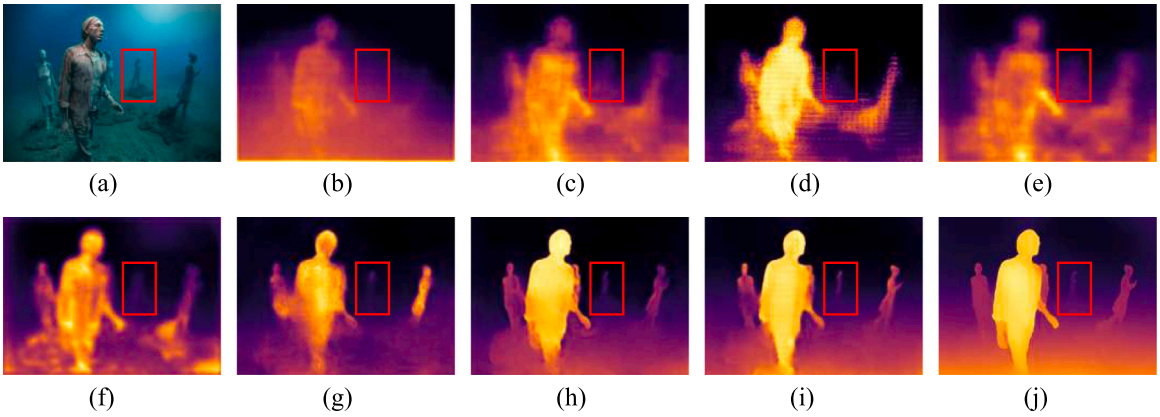

Fig. 4. Qualitative Comparison of the First Sample on the USOD10K Dataset [63]: (a) RGB; (b) IEbins [10]; (c) NeWCRFs [15]; (d) PixelFormer [11]; (e) URCDC-Depth [65]; (f) VA-DepthNet [66]; (g) UW-LapDepth [25]; (h) UDepth [21]; (i) CD-UDepth; (j) Ground Truth. CD-UDepth demonstrates superior depth estimation for distant objects.

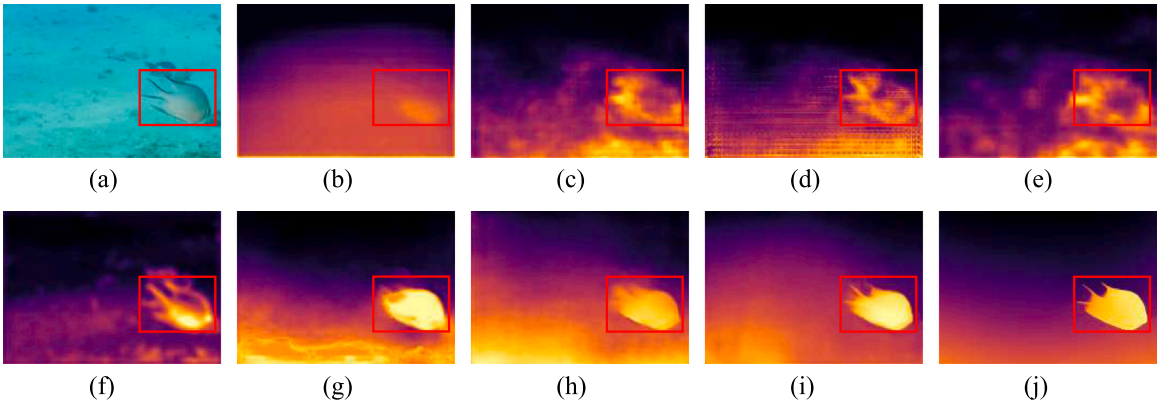

Fig. 5. Qualitative Comparison of the Second Sample on the USOD10K Dataset [63]: (a) RGB; (b) IEbins [10]; (c) NeWCRFs [15]; (d) PixelFormer [11]; (e) URCDC-Depth [65]; (f) VA-DepthNet [66]; (g) UW-LapDepth [25]; (h) UDepth [21]; (i) CD-UDepth; (j) Ground Truth. CD-UDepth achieves superior object edge estimation in blurry, predominantly blue-toned images.

CD-UDepth exhibits the lowest errors in the other metrics, demonstrating consistent performance across different error measurements.

Fig. 4 presents qualitative results for visual comparison. The first set of examples demonstrates the performance of CD-UDepth to distance attenuation and blurring. When background details are severely degraded, CD-UDepth can reliably extract effective structural cues and reconstruct objects in the background, significantly outperforming other methods. Fig. 5 demonstrates the performance of the proposed depth estimation algorithm when the image has a blue color cast and indistinct texture details, conditions typical in underwater environments. CD-UDepth provides the most accurate reconstruction of the fish in the image, preserving its shape and relative depth. Other algorithms show various limitations. NeWCRFs and PixelFormer, while able to locate the fish, cannot fully reconstruct its contours, and UW-LapDepth exhibits texture distortion in the predicted depth maps. These limitations in competing methods likely stem from their over-reliance on RGB information.

Fig. 6 demonstrates the performance of these depth estimation algorithms when the primary color tone of image is red. It can be observed that NeWCRFs, RCDC-Depth, and VA-DepthNet completely fail in estimating the depth of the fish in the image. This indicates that these algorithms rely too heavily on color cues, and when the overall color tone differs from typical underwater scenes, these algorithms exhibit poor generalization.

Fig. 7 presents a challenging case where our method reveals its limitations. In this case of a shell on an underwater sandy beach, CD-UDepth encounters difficulties in accurate depth reconstruction due to the poor visibility conditions. The low contrast nature of the scene reduces the effectiveness of both RGB and IMT spaces, while the uneven illumination distribution weakens the reliability of transmission-based

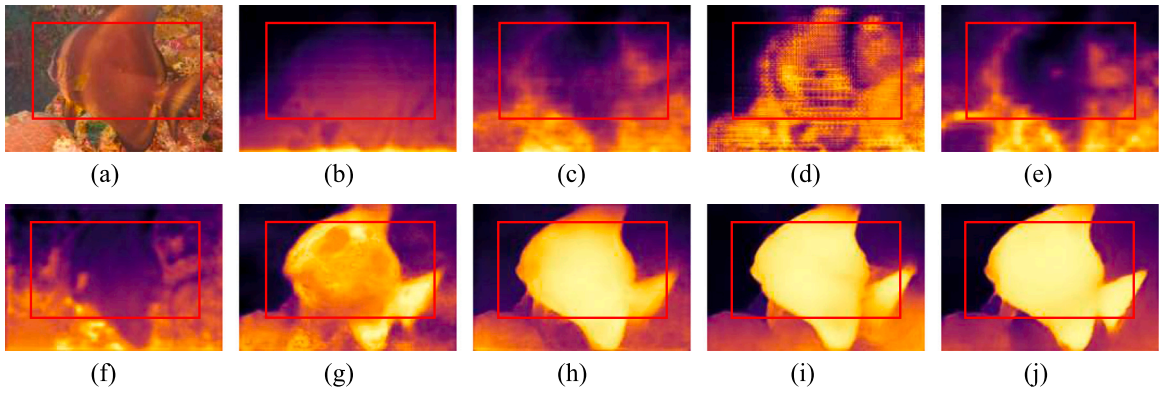

Fig. 6. Qualitative Comparison of the Third Sample on the USOD10K Dataset [63]: (a) RGB; (b) IEbins [10]; (c) NeWCRFs [15]; (d) PixelFormer [11]; (e) URCDC-Depth [65]; (f) VA-DepthNet [66]; (g) UW-LapDepth [25]; (h) UDepth [21]; (i) CD-UDepth; (j) Ground Truth. Most SOTA models exhibit significant misestimation in predominantly red-toned images.

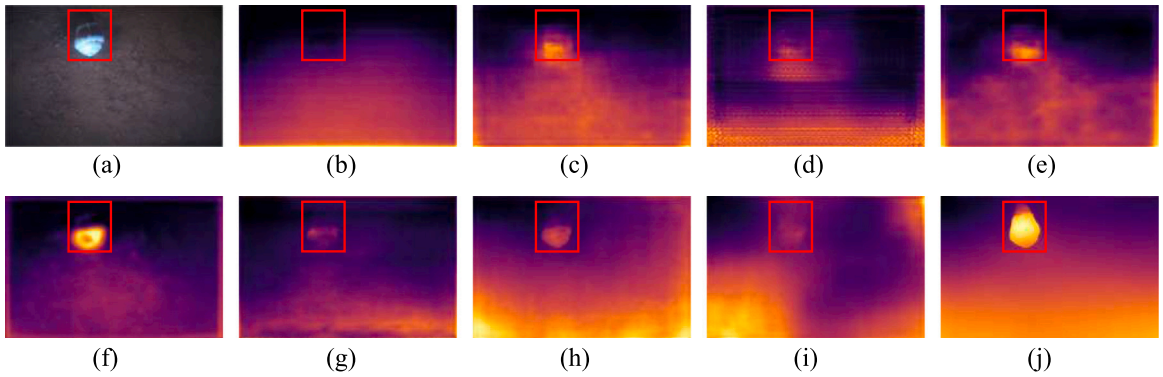

Fig. 7. Qualitative Comparison of the Fourth Sample on the USOD10K Dataset [63]: (a) RGB; (b) IEbins [10]; (c) NeWCRFs [15]; (d) PixelFormer [11]; (e) URCDC-Depth [65]; (f) VA-DepthNet [66]; (g) UW-LapDepth [25]; (h) UDepth [21]; (i) CD-UDepth; (j) Ground Truth. Depth estimation results of CD-UDepth showing challenges in scenes with uneven illumination and low contrast conditions.

Table 2
Qualitative comparison to SOTA methods on the FLSea dataset [64].

Method	δ_1 ↑	δ_2 ↑	δ_3 ↑	Abs Rel ↓	RMSE ↓	$\log_{10}$ ↓	Sq Rel ↓
IEbins [10]	0.375	0.667	0.841	0.554	0.159	0.173	0.139
NeWCRFs [15]	0.371	0.659	0.828	0.604	0.156	0.179	0.152
PixelFormer [11]	0.393	**0.693**	0.857	0.517	0.153	**0.164**	0.115
URCDC-Depth [65]	0.378	0.657	0.809	0.616	0.162	0.180	0.147
VA-DepthNet [66]	0.317	0.572	0.781	0.601	0.172	0.195	0.141
UW-LapDepth [25]	0.412	0.691	**0.859**	0.616	0.144	0.173	0.146
UDepth [21]	0.345	0.602	0.776	0.596	0.168	0.191	0.137
CD-UDepth	**0.396**	0.655	0.808	**0.486**	**0.141**	0.169	**0.100**

depth cues. These factors collectively affect the ability of confidence matrix to determine optimal fusion weights, leading to inconsistent depth estimation in both object and background regions.

Overall, IEbins and VA-DepthNet produce depth maps with blurred edges and low contrast, while PixelFormer's depth maps exhibit severe fence effects. Through the proposed qualitative comparison, CD-UDepth demonstrates superior adaptability and robustness in underwater environments compared to existing single-source models, particularly for distant blurred objects and areas with relatively little color and texture information. This advantage stems from the complementary fusion of features extracted from the color-light attenuation pattern.

Table 2 illustrates the performance of the models on the FLSea dataset [64]. The USOD50K dataset [63] features images with rich color variations and detailed object textures, leading to complex visual information. In contrast, the images in the FLSea dataset are characterized by more uniform and monotonous color tones. This characteristic has led to a notable enhancement in the performance of models on the FLSea dataset, with RMSE values generally falling between 0.15 and 0.16. CD-UDepth achieves the best performance in absolute error metrics, with Abs Rel of 0.486, RMSE of 0.141, and Sq Rel of 0.100, demonstrating its superior overall depth prediction accuracy. Notably, the model does not achieve optimal performance in accuracy metrics δ_2 and δ_3, indicating room for improvement in prediction stability across local regions. This is due to the annotation scheme of FLSea dataset where distant background regions are marked with zero values, and the strength of CD-UDepth in balanced foreground–background prediction becomes a limitation under such special annotation.

Fig. 8 illustrates the performance of these methods on the FLSea dataset. Compared to USOD10K, FLSea images exhibit a more monochromatic tone and concentrated distribution. Due to distant background regions being masked with 0 (appearing as black in depth maps), the depth maps obtained by depth estimation algorithms appear opposite to the ground truth values in background regions. During model training, we set maximum and minimum depth values and mask areas beyond these ranges to exclude extreme values that could affect the quantitative results. However, we chose to present the raw depth

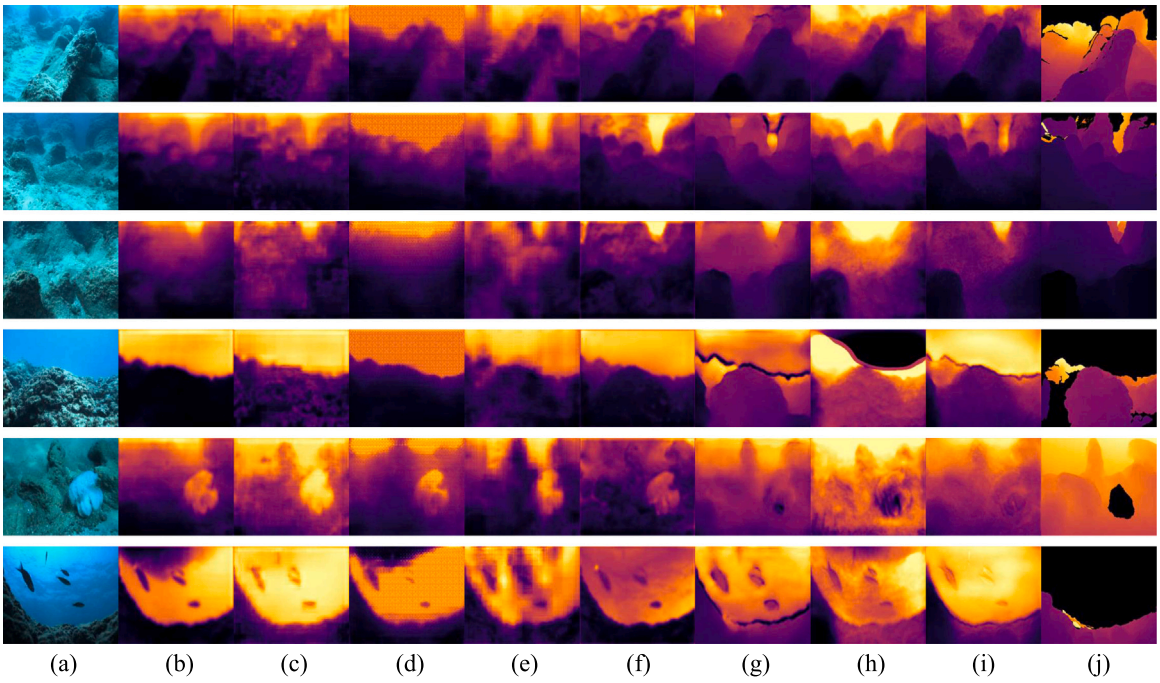

Fig. 8. Qualitative comparison on the Flsea dataset [64]: (a) RGB; (b) IEbins [10]; (c) NeWCRFs [15]; (d) PixelFormer [11]; (e) URCDC-Depth [65]; (f) VA-DepthNet [66]; (g) UW-LapDepth [25]; (h) UDepth [21]; (i) CD-UDepth; (j) Ground Truth.

predictions here without masking to better compare different methods and understand their ability to capture the overall scene structure. As shown in the results, even excluding the masked background regions, CD-UDepth demonstrates clear advantages in depth reconstruction of foreground areas, as particularly evident in the fourth set of examples. Notably, PixelFormer continues to exhibit significant gridding artifacts, while URCDC-Depth presents issues with low resolution.

4.5. Ablation study

To analyze the contributions of two depth information sources, RGB and IMT, to underwater depth estimation, Fig. 9 displays the depth maps generated by DESM-RGB and DESM-IMT, the confidence matrices, as well as the ultimate outcomes produced by CD-UDepth. In the first case, DESM-RGB achieves a more refined reconstruction of the closest statues, whereas DESM-IMT better represents the more distant background statues, aligning with their respective ranges of applicability: RGB assesses the foreground based on local color variations, while IMT relies on global light intensity changes to infer the background. Similar patterns can be observed in the other samples. The confidence matrices reasonably reflect their relative performance in color-distinct and background areas, with the automatic distribution of weights highlighting the advantages of the CDIFM. The final depth map, which takes into account both structure and detail, validates the framework's effective fusion, enhancing its robustness.

When comparing the single-source methods DESM-RGB and DESM-IMT, it was observed that they presented similar results in terms of RMSE and $\log_{10}$, as shown in Table 3. However, the differences in Abs Rel and Sq Rel indicators unveiled the limitations of a single information source in depth estimation tasks, while also hinting at the potential for complementary between different information sources. Compared with single-source models, the proposed CD-UDepth achieves significantly improved performance, reducing the RMSE by 16%.

In the detailed analysis of the CD-Udepth, a trend is noticed where error metrics seemed to decrease and then increase as the threshold θ rose. It is noteworthy that, when the parameter δ is set to 0.7, the model's performance reaches its peak, with the Abs Rel decreasing to 0.536 and the RMSE dropping by as much as 6.6%, compared to a direct concatenation without depth complementarity ($\delta = 0.0$). This finding underscores the significant impact of complementary depth in the fusion of two information sources on model performance, suggesting that either excessively high or low δ values are detrimental.

To investigate the effectiveness of different loss terms in our CD-UDepth, we conduct ablation experiments on the loss function components, as shown in Table 4. The experimental results reveal that all three combinations maintain relatively stable performance in terms of δ metrics, which evaluate the relative depth relationships. Specifically, the values of δ_1, δ_2, and δ_3 show minor variations across different loss combinations. However, significant differences are observed in error metrics. The incorporation of MSE loss substantially reduces the Abs Rel from 0.644 to 0.590, while the addition of Chamfer loss further decreases it to 0.536. Similar improvements are also reflected in RMSE and Sq Rel metrics. These results indicate that while the fundamental relative depth relationships are well preserved by the basic SILog loss, the additional loss terms primarily contribute to reducing absolute prediction errors and improving overall depth estimation accuracy.

4.6. Robustness study

In this Section, we present a comprehensive evaluation of CD-UDepth's robustness across two key dimensions: cross-region performance and contrast variation. These studies aim to assess the ability of the model to generalize across different underwater environments and its resilience to varying image quality conditions.

4.6.1. Cross-region robustness

To analyze the robustness of CD-UDepth across different marine regions, we tested the model trained in the Canyon region on the Red Sea region of the FLSea dataset. Table 5 showcases the performance comparison across three distinct underwater scenes: Big Dice Loop, Coral Table Loop, and Sub Pier. In the Big Dice Loop scene, CD-UDepth reduced the Abs Rel metric by 7.7% compared to its closest competitor, VA-DepthNet. CD-UDepth also excelled in RMSE and $\log_{10}$ metrics, reducing errors by 10.5% and 5.7% respectively. In the Coral Table Loop scene, while UW-LapDepth achieved the best accuracy metrics

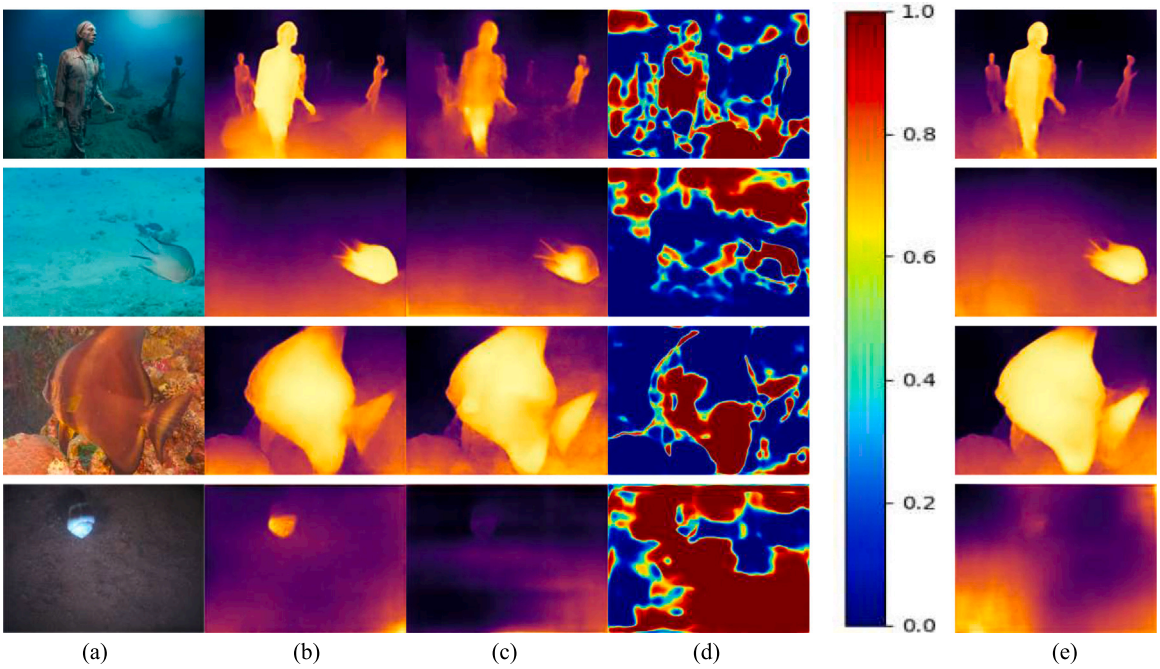

Fig. 9. Ablation study of dual-source DESMs and CD-UDepth: (a) RGB; (B) DESM-RGB; (c) DESM-IMT; (d) Confidence Map; (e) CD-UDepth.

Table 3
Ablation study of dual-source DESMs and CD-UDepth for different θ on the USOD10K dataset [63].

Model	θ	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	Abs Rel ↓	RMSE ↓	$\log_{10}$ ↓	Sq Rel ↓
DESM-RGB	–	0.444	0.669	0.790	0.766	0.147	0.191	0.159
DESM-IMT	–	0.455	0.672	0.793	0.668	0.146	0.184	0.140
CD-Udepth	0.0	0.468	0.693	0.814	0.550	0.136	0.174	0.097
	0.1	0.475	0.701	0.822	0.586	0.134	0.171	0.102
	0.2	0.473	0.701	0.821	0.562	0.133	0.170	0.091
	0.3	0.479	0.707	0.827	0.523	0.131	0.169	**0.083**
	0.4	0.483	0.710	0.827	0.538	0.130	0.169	0.086
	0.5	0.484	0.710	0.826	0.554	0.131	0.168	0.092
	0.6	**0.497**	0.717	0.831	0.537	0.128	**0.164**	0.084
	0.7	0.496	**0.720**	**0.833**	0.536	**0.127**	0.165	0.086
	0.8	0.484	0.710	0.827	**0.526**	0.131	0.168	0.088
	0.9	0.490	0.714	0.829	0.579	0.129	0.167	0.093

Table 4
Ablation study on different loss functions of CD-UDepth.

Loss function	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	Abs Rel ↓	RMSE ↓	$\log_{10}$ ↓	Sq Rel ↓
L_{SILog}	0.479	0.711	0.831	0.644	0.132	0.170	0.104
$L_{\text{SILog}} + L_{\text{MSE}}$	0.486	0.711	0.821	0.590	0.129	0.173	0.133
$L_{\text{SILog}} + L_{\text{MSE}} + L_{\text{Chamfer}}$	**0.496**	**0.720**	**0.833**	**0.536**	**0.127**	**0.165**	**0.086**

and NeWCRFs obtained the lowest Abs Rel, CD-UDepth maintained competitive performance with the lowest RMSE of 0.128. In the Sub Pier scene, CD-UDepth maintained its advantage, particularly in the RMSE metric, which is 3.7% lower than the second-best IEbins.

Overall, CD-UDepth exhibited superior performance across all three scenes, verifying its robustness and adaptability in diverse underwater environments. It is noteworthy that some methods showed significant performance variations across different scenes. For instance, NeWCRFs performed relatively well in the Coral Table Loop scene but poorly in the Sub Pier scene, with its Abs Rel increasing from 0.431 to 0.540, a 25.3% increase. This underscores the importance of developing algorithms capable of maintaining stable performance across various underwater environments.

4.6.2. Contrast robustness

To further investigate the robustness of the CD-UDepth model in underwater environments, we simulate images with different levels of low contrast based on the benchmark proposed by [67]. Let I and G be the original image and processed image, the contrast adjustment formula is as follows:

$$G = (I - \bar{I}) \cdot c + \bar{I}, \tag{15}$$

in which $\bar{I}$ is the mean value of I, and c is a positive adjustment factor to control the contrast of images. The contrast of an image decreases as c decreases.

Table 6 shows the performance of the single-source models DESM-RGB and DESM-IMT as well as the CD-UDepth model when subjected to simulated images with lowered contrast levels. Notably, at $c = 1$, where the images are unadjusted for contrast, the results function as a baseline for comparative analysis. The CD-UDepth consistently surpasses both DESM-RGB and DESM-IMT in performance under various degrees of low contrast, indicating its superior adaptability to challenging visual conditions.

Upon examining the comparative metrics between the two single-source input spaces, DESM-RGB has a slight edge over DESM-IMT

Table 5
Cross-region robustness study.

Scene	Method	δ_1 ↑	δ_2 ↑	δ_3 ↑	Abs Rel ↓	RMSE ↓	$\log_{10}$ ↓	Sq Rel ↓
Big Dice Loop	IEbins [10]	0.178	0.344	0.495	1.234	0.188	0.299	0.324
	NeWCRFs [15]	0.079	0.198	0.366	1.427	0.206	0.351	0.352
	PixelFormer [11]	0.168	0.364	0.532	1.163	0.170	0.289	0.270
	URCDC-Depth [65]	0.123	0.267	0.401	1.497	0.206	0.347	0.376
	VA-DepthNet [66]	0.166	0.334	0.528	0.933	0.153	0.262	0.180
	UW-LapDepth [25]	0.102	0.234	0.383	1.433	0.204	0.345	0.350
	UDepth [21]	0.101	0.240	0.422	1.370	0.193	0.333	0.328
	CD-Udepth	**0.194**	**0.398**	**0.617**	**0.861**	**0.137**	**0.247**	**0.143**
Coral Table Loop	IEbins [10]	0.375	0.666	0.860	0.599	0.142	0.162	0.173
	NeWCRFs [15]	0.392	0.680	0.877	**0.431**	0.146	0.155	**0.092**
	PixelFormer [11]	0.344	0.618	0.819	0.651	0.155	0.178	0.164
	URCDC-Depth [65]	0.291	0.582	0.809	0.685	0.149	0.186	0.164
	VA-DepthNet [66]	0.416	0.726	0.897	0.433	0.134	0.146	0.104
	UW-LapDepth [25]	**0.481**	**0.769**	**0.921**	0.467	0.131	**0.132**	0.100
	UDepth [21]	0.296	0.583	0.801	0.551	0.162	0.187	0.125
	CD-Udepth	0.437	0.716	0.848	0.554	**0.128**	0.150	0.146
Sub Pier	IEbins [10]	**0.341**	**0.635**	0.836	0.574	0.107	0.177	0.094
	NeWCRFs [15]	0.323	0.599	0.780	**0.540**	0.125	0.195	**0.079**
	PixelFormer [11]	0.324	0.615	0.808	0.652	0.110	0.184	0.108
	URCDC-Depth [65]	0.127	0.328	0.597	1.045	0.156	0.273	0.203
	VA-DepthNet [66]	0.113	0.338	0.624	0.964	0.151	0.262	0.182
	UW-LapDepth [25]	0.206	0.481	0.683	0.887	0.128	0.233	0.164
	UDepth [21]	0.248	0.508	0.739	0.674	0.162	0.205	0.147
	CD-Udepth	0.339	0.631	**0.842**	0.556	**0.103**	**0.172**	0.085

Table 6
Robustness study at varying levels of contrast on the USOD10K dataset [63].

Model	c	δ_1 ↑	δ_2 ↑	δ_3 ↑	Abs Rel ↓	RMSE ↓	$\log_{10}$ ↓	Sq Rel ↓
DESM-RGB	0.3	0.390	0.609	0.737	1.246	0.177	0.226	0.337
	0.4	0.406	0.628	0.756	1.084	0.167	0.214	0.267
	0.5	0.418	0.642	0.771	0.970	0.159	0.204	0.223
	0.75	0.440	0.662	0.786	0.827	0.149	0.193	0.175
	1.0[a]	**0.444**	**0.669**	**0.790**	**0.766**	**0.147**	**0.191**	**0.159**
DESM-IMT	0.3	0.388	0.603	0.739	0.977	0.178	0.219	0.249
	0.4	0.409	0.629	0.758	0.875	0.168	0.207	0.215
	0.5	0.429	0.652	0.779	0.786	0.158	0.196	0.188
	0.75	0.450	0.669	0.792	0.688	0.148	0.186	0.147
	1.0[a]	**0.455**	**0.672**	**0.793**	**0.668**	**0.146**	**0.184**	**0.140**
CD-UDepth	0.3	0.431	0.654	0.785	0.717	0.155	0.195	0.151
	0.4	0.438	0.658	0.786	0.667	0.150	0.191	0.129
	0.5	0.458	0.683	0.805	0.624	0.141	0.182	0.115
	0.75	0.479	0.703	0.823	0.577	0.132	0.171	0.099
	1.0[a]	**0.496**	**0.720**	**0.833**	**0.536**	**0.127**	**0.165**	**0.086**

[a] When $c = 1.0$, there is no contrast adjustment performed on the underwater image.

in terms of Abs Rel, suggesting that it can maintain relative depth accuracy more effectively. Conversely, DESM-IMT outperforms DESM-RGB on RMSE metric, implying that it is better at estimating depth consistently across the entire image.

The table further reveals the models' resilience to contrast fluctuations, by showing the metric variations from the lowest contrast setting to normal conditions. DESM-RGB demonstrates minimal fluctuation in both Abs Rel and Sq Rel, while the CD-UDepth model shows stability in the face of contrast variations in terms of RMSE and $\log_{10}$, with the smallest deviations being 0.028 and 0.03, respectively. While all models exhibit some level of performance degradation with decreased contrast, CD-UDepth's design can offer the most consistent performance across different contrast levels.

4.7. Complexity evaluation

We further analyze the computational complexity of different methods in terms of parameters, memory consumption, computational cost (FLOPs), and inference time, as shown in Table 7. The comparison reveals that most transformer-based methods, such as PixelFormer and VA-DepthNet, have relatively high computational requirements, with parameters exceeding 250M and memory consumption over 1000 MB.

Table 7
Complexity comparison of different underwater depth estimation methods.

Method	Params (M)	Memory (MB)	Flops (G)	Inference time (ms)
IEbins [10]	272.8	1091.2	622.9	178.3
NeWCRFs [15]	270.3	1081.3	280.8	102.8
PixelFormer [11]	258.3	1033.0	273.3	81.1
URCDC-Depth [65]	333.6	1334.6	416.0	133.1
VA-DepthNet [66]	257.1	1028.3	382.5	158.8
UPGformer [55]	15.81	63.2	37.4	–
UW-LapDepth [25]	58.0	232.1	91.1	61.2
UDepth [21]	**15.7**	**62.7**	**37.4**	**20.0**
CD-Udepth	70.9	283.6	184.7	88.6

Table 8
Inference time of CD-UDepth under different input resolutions.

Input resolution	Standard (640 × 480)	256 × 256	512 × 512	1024 × 1024
Time (ms)	88.6	64.5	86.6	170.2

Among all methods, URCDC-Depth has the highest complexity with 333.6M parameters and 1334.6 MB memory usage. In terms of inference speed, transformer-based methods generally exhibit longer

processing times, with IEbins requiring 178.3 ms per image and VA-DepthNet taking 158.8 ms. In contrast, lightweight models like UDepth and UPGformer demonstrate remarkable efficiency, requiring only around 15M parameters. Notably, UDepth achieves the fastest inference time of 20.0 ms while maintaining the lowest resource requirements. Our CD-UDepth achieves a balanced trade-off between model complexity and performance, with moderate resource requirements of 70.9M parameters and competitive inference speed of 88.6 ms while maintaining strong performance as demonstrated in previous experiments. Furthermore, we analyzed the inference time of CD-UDepth under different input resolutions, as shown in Table 8. The model achieves 86.6 ms at standard resolution (640 × 480), and scales well with different input sizes, ranging from 64.5 ms at 256 × 256 to 170.2 ms at 1024 × 1024.

5. Conclusion

This paper has proposed CD-UDepth, an innovative underwater depth estimation model that leverages dual information sources. The model introduces an IMT space guided by light attenuation, which exhibits enhanced depth perception for distant backgrounds in underwater images compared to the traditional RGB space. To harness the benefits of both color and light attenuation, we have designed a complementary dual-source information fusion module that uses confidence levels to guide the merging of information from both sources, producing the final depth map. Extensive experiments have demonstrated the superiority of CD-UDepth over existing methods. Notably, our model achieves a 30% reduction in Abs Rel compared to models that only utilize RGB space as an information source, showcasing greater robustness and accuracy in underwater depth estimation tasks.

While CD-UDepth demonstrates significant improvements in underwater depth estimation, there remain opportunities for further advancements. The integration of physical underwater imaging models could improve performance in extreme conditions. Further investigation is needed to optimize the computational efficiency for real-time applications and to evaluate the generalization capability of the model across different marine environments. These developments would enhance the practical applicability of underwater depth estimation in marine robotics and underwater scene understanding. Additionally, the proposed underwater depth estimation framework provides insights for other computer vision tasks in challenging environments. The dual-source fusion strategy and confidence-guided mechanism could be extended to other domains where traditional RGB-based methods face similar degradation issues, such as depth estimation in fog, smoke, or low-light conditions.

CRediT authorship contribution statement

Jiawei Guo: Writing – original draft, Software, Methodology. **Jieming Ma:** Writing – review & editing, Supervision, Conceptualization. **Feiyang Sun:** Validation, Software, Methodology. **Zhiqiang Gao:** Investigation. **Ángel F. García-Fernández:** Writing – original draft, Supervision. **Hai-Ning Liang:** Writing – review & editing. **Xiaohui Zhu:** Writing – review & editing, Conceptualization. **Weiping Ding:** Project administration.

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 research is supported by the Natural Science Foundation of China (Grant No. 62472361), the Suzhou Science and Technology Project-Key Industrial Technology Innovation (SYG202122), 2024 Suzhou Innovation Consortium Construction Project, the XJTLU Postgraduate Research Scholarship (Grand No. PGRS1906004), the XJTLU AI University Research Centre, Zooming New Energy-XJTLU Smart Energy Joint Laboratory, Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU and SIP AI innovation platform (YZCXPT2022103).

Data availability

The implementation of the proposed work will be available at: https://github.com/cainsmile/CD-UDepth.

References

[1] S.T. Digumarti, G. Chaurasia, A. Taneja, R. Siegwart, A. Thomas, P. Beardsley, Underwater 3D capture using a low-cost commercial depth camera, in: 2016 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2016, pp. 1–9.
[2] P.N. Leite, A.M. Pinto, Fusing heterogeneous tri-dimensional information for reconstructing submerged structures in harsh sub-sea environments, Inf. Fusion 103 (2024) 102126.
[3] Y. Yang, Y. Xiao, T. Li, A survey of autonomous underwater vehicle formation: Performance, formation control, and communication capability, IEEE Commun. Surv. Tutor. 23 (2) (2021) 815–841.
[4] B. Zhang, D. Ji, S. Liu, X. Zhu, W. Xu, Autonomous underwater vehicle navigation: A review, Ocean Eng. (2023) 113861.
[5] D. McLeod, J. Jacobson, M. Hardy, C. Embry, Autonomous inspection using an underwater 3D LiDAR, in: 2013 OCEANS-San Diego, IEEE, 2013, pp. 1–8.
[6] H. Lu, Y. Zhang, Y. Li, Q. Zhou, R. Tadoh, T. Uemura, H. Kim, S. Serikawa, Depth map reconstruction for underwater Kinect camera using inpainting and local image mode filtering, IEEE Access 5 (2017) 7115–7122.
[7] A.A. Sheikh, E. Felemban, M. Felemban, S.B. Qaisar, Challenges and opportunities for underwater sensor networks, in: 2016 12th International Conference on Innovations in Information Technology, IIT, IEEE, 2016, pp. 1–6.
[8] K.M. Awan, P.A. Shah, K. Iqbal, S. Gillani, W. Ahmad, Y. Nam, et al., Underwater wireless sensor networks: A review of recent issues and challenges, Wirel. Commun. Mob. Comput. 2019 (2019).
[9] K. Sun, W. Cui, C. Chen, Review of underwater sensing technologies and applications, Sensors 21 (23) (2021) 7849.
[10] S. Shao, Z. Pei, X. Wu, Z. Liu, W. Chen, Z. Li, IEBins: Iterative elastic bins for monocular depth estimation, 2023, arXiv preprint arXiv:2309.14137.
[11] A. Agarwal, C. Arora, Attention attention everywhere: Monocular depth prediction with skip attention, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5861–5870.
[12] Z. Li, X. Wang, X. Liu, J. Jiang, Binsformer: Revisiting adaptive bins for monocular depth estimation, 2022, arXiv preprint arXiv:2204.00987.
[13] Z. Li, Z. Chen, X. Liu, J. Jiang, Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation, Mach. Intell. Res. 20 (6) (2023) 837–854.
[14] S.F. Bhat, I. Alhashim, P. Wonka, Adabins: Depth estimation using adaptive bins, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4009–4018.
[15] W. Yuan, X. Gu, Z. Dai, S. Zhu, P. Tan, Neural window fully-connected crfs for monocular depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3916–3925.
[16] J. Jun, J.-H. Lee, C. Lee, C.-S. Kim, Depth map decomposition for monocular depth estimation, in: European Conference on Computer Vision, Springer, 2022, pp. 18–34.
[17] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The kitti dataset, Int. J. Robot. Res. 32 (11) (2013) 1231–1237.
[18] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12, Springer, 2012, pp. 746–760.
[19] H. Gupta, K. Mitra, Unsupervised single image underwater depth estimation, in: 2019 IEEE International Conference on Image Processing, ICIP, IEEE, 2019, pp. 624–628.
[20] 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.

[21] B. Yu, J. Wu, M.J. Islam, Udepth: Fast monocular depth estimation for visually-guided underwater robots, in: 2023 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2023, pp. 3116–3123.

[22] M. Yang, J. Hu, C. Li, G. Rohde, Y. Du, K. Hu, An in-depth survey of underwater image enhancement and restoration, IEEE Access 7 (2019) 123638–123657.

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

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

[25] W. Chen, X. Luo, F. Li, D. Wang, Estimation of underwater monocular depth using Laplacian pyramid-based depth residuals, in: 2023 13th International Conference on Information Science and Technology, ICIST, IEEE, 2023, pp. 40–47.

[26] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.

[27] S.F. Bhat, I. Alhashim, P. Wonka, Localbins: Improving depth estimation by learning local distributions, in: European Conference on Computer Vision, Springer, 2022, pp. 480–496.

[28] Z. Li, X. Liu, N. Drenkow, A. Ding, F.X. Creighton, R.H. Taylor, M. Unberath, Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6197–6206.

[29] A. Gordon, H. Li, R. Jonschkowski, A. Angelova, Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8977–8986.

[30] J. Kopf, X. Rong, J.-B. Huang, Robust consistent video depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1611–1621.

[31] W. Zhou, E. Zhou, G. Liu, L. Lin, A. Lumsdaine, Unsupervised monocular depth estimation from light field image, IEEE Trans. Image Process. 29 (2019) 1606–1617.

[32] H. Zhan, R. Garg, C.S. Weerasekera, K. Li, H. Agarwal, I. Reid, Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 340–349.

[33] X. Wang, H. Luo, Z. Wang, J. Zheng, X. Bai, Self-supervised multi-frame depth estimation with visual-inertial pose transformer and monocular guidance, Inf. Fusion 108 (2024) 102363.

[34] H.-X. Chen, K. Li, Z. Fu, M. Liu, Z. Chen, Y. Guo, Distortion-aware monocular depth estimation for omnidirectional images, IEEE Signal Process. Lett. 28 (2021) 334–338.

[35] H. Hu, B. Yang, Z. Qiao, S. Liu, J. Zhu, Z. Liu, W. Ding, D. Zhao, H. Wang, SeasonDepth: Cross-season monocular depth prediction dataset and benchmark under multiple environments, in: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2023, pp. 11384–11389.

[36] L. Kong, S. Xie, H. Hu, L.X. Ng, B. Cottereau, W.T. Ooi, Robodepth: Robust out-of-distribution depth estimation under corruptions, Adv. Neural Inf. Process. Syst. 36 (2024).

[37] C. Long, W. Zhang, Z. Chen, H. Wang, Y. Liu, P. Tong, Z. Cao, Z. Dong, B. Yang, SparseDC: Depth completion from sparse and non-uniform inputs, Inf. Fusion 110 (2024) 102470.

[38] J. Zhou, T. Yang, W. Zhang, Underwater vision enhancement technologies: A comprehensive review, challenges, and recent trends, Appl. Intell. 53 (3) (2023) 3594–3621.

[39] D. Zhang, C. Wu, J. Zhou, W. Zhang, Z. Lin, K. Polat, F. Alenezi, Robust underwater image enhancement with cascaded multi-level sub-networks and triple attention mechanism, Neural Netw. 169 (2024) 685–697.

[40] F.A. Dharejo, I.I. Ganapathi, M. Zawish, B. Alawode, M. Alathbah, N. Werghi, S. Javed, SwinWave-SR: Multi-scale lightweight underwater image super-resolution, Inf. Fusion 103 (2024) 102127.

[41] J. Zhou, S. Wang, D. Zhang, Q. Jiang, K. Jiang, Y. Lin, Decoupled variational retinex for reconstruction and fusion of underwater shallow depth-of-field image with parallax and moving objects, Inf. Fusion (2024) 102494.

[42] S. An, L. Xu, Z. Deng, H. Zhang, DNIM: Deep-sea netting intelligent enhancement and exposure monitoring using bio-vision, Inf. Fusion 113 (2025) 102629.

[43] S. Fayaz, S.A. Parah, G. Qureshi, Underwater object detection: Architectures and algorithms–a comprehensive review, Multimedia Tools Appl. 81 (15) (2022) 20871–20916.

[44] Y. Tian, M. Khishe, R. Karimi, E. Hashemzadeh, O. Pakdel Azar, Underwater image detection and recognition using radial basis function neural networks and chimp optimization algorithm, Circuits Systems Signal Process. 42 (7) (2023) 3963–3982.

[45] F. Yu, B. He, J.-X. Liu, Underwater targets recognition based on multiple AUVs cooperative via recurrent transfer-adaptive learning (RTAL), IEEE Trans. Veh. Technol. 72 (2) (2022) 1574–1585.

[46] L. Christensen, J. de Gea Fernández, M. Hildebrandt, C.E.S. Koch, B. Wehbe, Recent advances in ai for navigation and control of underwater robots, Curr. Robot. Rep. 3 (4) (2022) 165–175.

[47] A.M. Pinto, A.C. Matos, MARESye: A hybrid imaging system for underwater robotic applications, Inf. Fusion 55 (2020) 16–29.

[48] D.Q. Huy, N. Sadjoli, A.B. Azam, B. Elhadidi, Y. Cai, G. Seet, Object perception in underwater environments: A survey on sensors and sensing methodologies, Ocean Eng. 267 (2023) 113202.

[49] S. Rahman, A. Quattrini Li, I. Rekleitis, SVIn2: A multi-sensor fusion-based underwater SLAM system, Int. J. Robot. Res. 41 (11–12) (2022) 1022–1042.

[50] 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.

[51] K. He, J. Sun, X. Tang, Single image haze removal using dark channel prior, IEEE Trans. Pattern Anal. Mach. Intell. 33 (12) (2011) 2341–2353, http://dx.doi.org/10.1109/TPAMI.2010.168.

[52] A. Galdran, D. Pardo, A. Picón, A. Alvarez-Gila, Automatic red-channel underwater image restoration, J. Vis. Commun. Image Represent. 26 (2015) 132–145.

[53] 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, IEEE, 2015, pp. 4952–4956.

[54] D. Berman, D. Levy, S. Avidan, T. Treibitz, Underwater single image color restoration using haze-lines and a new quantitative dataset, IEEE Trans. Pattern Anal. Mach. Intell. 43 (8) (2020) 2822–2837.

[55] C. Wang, H. Xu, G. Jiang, M. Yu, T. Luo, Y. Chen, Underwater monocular depth estimation based on physical-guided transformer, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–16.

[56] C.O. Ancuti, C. Ancuti, C. De Vleeschouwer, P. Bekaert, Color balance and fusion for underwater image enhancement, IEEE Trans. Image Process. 27 (1) (2017) 379–393.

[57] 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: Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part I 19, Springer, 2018, pp. 678–688.

[58] L. Zhang, B. He, Y. Song, T. Yan, Underwater image feature extraction and matching based on visual saliency detection, in: OCEANS 2016 - Shanghai, 2016, pp. 1–4.

[59] Y. Wang, N. Li, Z. Li, Z. Gu, H. Zheng, B. Zheng, M. Sun, An imaging-inspired no-reference underwater color image quality assessment metric, Comput. Electr. Eng. 70 (2018) 904–913.

[60] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[61] A. Mertan, D.J. Duff, G. Unal, Single image depth estimation: An overview, Digit. Signal Process. 123 (2022) 103441.

[62] H. Fan, H. Su, L.J. Guibas, A point set generation network for 3d object reconstruction from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 605–613.

[63] L. Hong, X. Wang, G. Zhang, M. Zhao, USOD10K: A new benchmark dataset for underwater salient object detection, IEEE Trans. Image Process. (2023) http://dx.doi.org/10.1109/TIP.2023.3266163, 1–1.

[64] Y. Randall, FLSea: Underwater Visual-Inertial and Stereo-Vision Forward-Looking Datasets (Ph.D. thesis), University of Haifa (Israel), 2023.

[65] S. Shao, Z. Pei, W. Chen, R. Li, Z. Liu, Z. Li, URCDC-Depth: Uncertainty Rectified Cross-Distillation with CutFlip for Monocular Depth Estimation, 2023, arXiv preprint arXiv:2302.08149.

[66] C. Liu, S. Kumar, S. Gu, R. Timofte, L. Van Gool, Va-depthnet: A variational approach to single image depth prediction, 2023, arXiv preprint arXiv:2302.06556.

[67] D. Hendrycks, T. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, 2019, arXiv preprint arXiv:1903.12261.