

Underwater Monocular Depth Estimation: A Survey of Theory, Methods, and Applications

Peixian Zhuang^{1,*}, Qi Tian^{1,*}, Yijian Wang^{2,*}, Jiaheng Han¹, and Yuchun Wang¹

ABSTRACT

Underwater monocular depth estimation aims to infer the distances between scene objects from the camera from a single underwater image. Nevertheless, the propagation of visible light in water is subject to wavelength-dependent absorption and scattering, which introduce color distortion, low contrast, blur, and texture degradation, thereby hindering the reliable extraction of depth cues. The degradation of underwater images varies with the propagation distance of visible light, providing a physical basis for scene distance estimation. Leveraging this property, underwater monocular depth estimation has evolved from inferring distances using single physical cues to comprehensive estimation approaches that integrate deep learning with multi-source observations. Starting from the theoretical foundations of underwater imaging and light propagation, Bayesian estimation, and prior constraints, our survey classifies existing methods into three categories: traditional physical-prior methods, deep-learning-driven methods, and multimodal fusion methods. Moreover, we review the utilization of depth cues across a range of underwater tasks, including underwater image enhancement and restoration, underwater three-dimensional (3D) reconstruction, underwater object detection, autonomous underwater vehicle (AUV) navigation and localization, and underwater robotic operations. On this basis, our survey analyzes current challenges, including complex optical degradation, insufficient cross-domain generalization, the scarcity of high-quality real-world data, and the difficulty of balancing accuracy with computational efficiency. Further, we outline promising avenues for future research, including the integration of underwater imaging priors and deep learning, the improvement of multimodal fusion reliability, the development and benchmarking of real-world underwater depth datasets, and the advancement of real-time depth estimation tailored for underwater intelligent devices.

KEYWORDS

Underwater scene, monocular depth, formation model, deep learning, multimodal fusion.

1 Introduction

Depth is a key physical quantity for characterizing the three-dimensional (3D) geometry of underwater scenes and serves as an essential foundation for advancing underwater visual perception from two-dimensional (2D) appearance understanding to 3D spatial cognition. Conventional underwater images primarily record two-dimensional projection information such as color, texture, and semantics of scenes, whereas underwater depth estimation further recovers the distance from scene points relative to the camera and their spatial distribution from one or more images, thereby providing reliable geometric information for tasks such as underwater 3D reconstruction^[1], underwater object detection and localization^[2], autonomous underwater vehicle (AUV) navigation^[3], and underwater robotic operations^[4].

According to the modality of visual observation, underwater depth estimation can be primarily categorized into monocular depth estimation and stereo/multi-view depth estimation.

Monocular methods infer scene depth from visual cues such as color, texture, edges, perspective, and semantics within a single image. They feature simple system configurations, flexible deployment, and low hardware cost, making them more suitable for underwater platforms with constrained space, power, and payload capacity. But their estimation results generally suffer from scale ambiguity^[4], due to the absence of explicit constraints such as disparity and geometric baselines^[5]. Stereo and multi-view methods recover scene depth by leveraging inter-view correspondences, disparity, and camera intrinsic and extrinsic parameters. Under the conditions of accurate calibration and geometric modeling^[6], they can produce metric depth with actual physical meaning, but their performance heavily depends on cross-view matching, camera calibration^[7], and multi-view consistency. According to output scale, underwater depth can also be divided into relative depth and metric depth: relative depth primarily describes the depth ordering and spatial hierarchy among different scene regions, whereas metric depth reports the distance between a

¹  Mariner Lab, Key Laboratory of Knowledge Automation for Industrial Processes, Ministry of Education, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China. E-mail: zhuangpeixian0624@163.com

² School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

scene point and the camera in actual physical units such as centimeters or meters.

Compared with terrestrial scenes imaged through air, visible light propagation in underwater environments is considerably more complex. Visible light undergoes absorption and scattering as it travels through water, with significant variations in attenuation across different wavelengths. Water turbidity, suspended particles, artificial illumination, and nonuniform lighting further cause image degradation phenomena such as color shifts, contrast reduction, edge blurring, and texture attenuation [8, 9, 10]. These degradations can weaken or even distort the color, texture, edge, and semantic cues used by monocular depth estimation, and reduce the reliability of correspondence matching between stereo or multi-view images. In addition, underwater cameras typically observe external scenes through waterproof housings, where light refracts at the interfaces between different media such as air, waterproof glass, and water [7], introducing deviations into the conventional pinhole imaging model and epipolar geometry [6], which further complicates camera calibration, disparity estimation, and metric-scale recovery. Consequently, depth-estimation methods designed for terrestrial imagery cannot be directly adapted to complex and dynamic underwater environments; instead, specialized depth estimation approaches must be developed by integrating underwater light propagation mechanisms, refractive geometric models, and specific characteristics of underwater data distributions.

Existing research on underwater depth estimation has largely evolved along three technical lines: traditional physical priors, deep-learning-driven estimation, and multimodal fusion. Traditional physical-prior methods establish physical relationships between underwater light-propagation distance and image observations, and estimate scene depth by exploiting priors such as dark channels, color differences, light attenuation [11], image blurriness [12], and backscattering [13]. For instance, some approaches first classify the water type and then estimate the relative depth of underwater images using corresponding light attenuation parameters [14]. Such methods exhibit strong physical interpretability and do not rely on large-scale training data, but they are typically built upon assumptions such as specific water types, uniform illumination, or fixed imaging parameters, which limits their applicability in complex and dynamic environments.

With the advancement of deep learning, Convolutional Neural Networks (CNNs) [15], Transformers [16, 17], and Mamba architectures based on state-space models [18] have been increasingly applied to underwater depth estimation. These methods can jointly learn multi-level depth cues, involving color, texture, edges, semantics, and global structure, from data, thereby reducing reliance on single hand-crafted priors and improving feature representation capabilities in complex scenes. However, deep learning-based methods generally require large-scale underwater data with accurate depth annotations [19]. In practice, acquiring real underwater depth data is difficult and annotation costs are high [20]. Moreover, significant domain gaps exist across different water bodies, lighting conditions, and imaging devices [8, 21, 22], which makes models vulnerable to issues such as poor generalization, scale drift, and difficulty in quantifying uncertainty.

To compensate for the limitations of purely visual observations under low-visibility, weak-texture, and scale-

recovery scenarios, some studies have further introduced sonar, echo sounders [23, 24], and other ranging sensors [4] to fuse visual information with acoustic or range data. For instance, distance observations provided by single-beam echo sounders have been integrated into Simultaneous Localization and Mapping (SLAM) systems [5] to calibrate the scale ambiguity inherent in monocular vision. Multimodal fusion can combine the robustness of acoustic sensing in turbid water with the high spatial resolution of visual imagery, forming complementary advantages in absolute-scale recovery and complex environment perception. However, it faces challenges such as heterogeneous data representations, spatial extrinsic parameter calibration [23], temporal synchronization [24], and cross-modal feature alignment [25]. Underwater depth estimation has evolved from relative-depth inference relying on single optical physical cues to data-driven approaches using multi-level visual representations, and is further progressing toward the collaborative modeling of vision, acoustic, and other sensory information.

To date, existing surveys on depth estimation have primarily focused on monocular depth estimation in terrestrial scenes. They have systematically categorized deep learning-based monocular depth estimation methods [26] and discussed supervised and unsupervised learning, network architectures, datasets, and evaluation metrics. Such reviews have laid a relatively complete foundation for understanding the development of general monocular depth estimation, but they typically assume that images conform to imaging principles in air and rarely provide systematic analyses of how underwater absorption, scattering, wavelength-dependent attenuation, and medium refraction affect depth cues and geometric models. Furthermore, existing works often treat underwater depth estimation merely as a specific application within general depth estimation, underwater image enhancement and restoration, or underwater 3D vision, and there remains a lack of a systematic survey that starts from underwater imaging mechanisms and coherently covers theoretical foundations, core methods, and typical applications.

To this end, our survey presents a comprehensive review of the theoretical foundations, technical approaches, and application scenarios of underwater depth estimation, organizing related studies into a three-level framework of theory, methods, and applications, as illustrated in Fig. 1. At the theoretical level, we focus on introducing underwater image formation and light-propagation models, refractive geometry, Bayesian estimation, visual prior constraints, and deep feature representation mechanisms, elucidating the intrinsic relationships between underwater image observations and scene depth. At the methodological level, we systematically review underwater depth estimation methods based on traditional physical priors, deep-learning-driven approaches, and multimodal fusion, analyzing the fundamental assumptions, depth cues, scale recovery mechanisms, and applicable conditions of different methods. At the application level, we further summarize the roles of underwater depth information in image enhancement and restoration, 3D scene reconstruction, object detection, AUV navigation and localization, and precise underwater robotic operations. Finally, current challenges faced by underwater depth estimation concerning data, models, environmental adaptability, and practical deployment are analyzed, and potential future directions are discussed.

This survey primarily focuses on the studies of underwater depth estimation and its applications, together with relevant theoretical and methodological foundations. Representative approaches are selected based on depth cues, adaptations to underwater imaging challenges, and applications of depth information, covering early works and recent advances. The main contributions of our article are summarized as follows: (1) A unified theory–method–application survey framework is established for underwater depth estimation, systematically reviewing research in this field and forming a complete study pipeline from fundamental mechanisms to practical applications. (2) Starting from factors such as underwater light absorption and scattering, wavelength-selective attenuation, non-uniform illumination, and medium refraction, the coupling relationships among underwater image degradation, visual depth cues, and imaging geometry are analyzed, elucidating why terrestrial depth estimation methods deteriorate underwater. (3) Existing underwater depth estimation methods are categorized into three classes: traditional physical-prior-based methods, deep-learning-driven methods, and multimodal fusion methods. Their physical assumptions, supervision information, scale recovery capability, environmental robustness, computational complexity, and deployment requirements are compared, and the complementary relationships among different technical routes are analyzed. (4) The progress of underwater depth estimation in applications of underwater image enhancement and restoration, underwater 3D reconstruction, underwater object detection, AUV navigation and localization, and underwater robotic operations is systematically summarized. Additionally, the key challenges currently facing the field are distilled, and future research areas including data construction, fusion of physical models and data-driven methods, cross-modal perception, and lightweight deployment are provided.

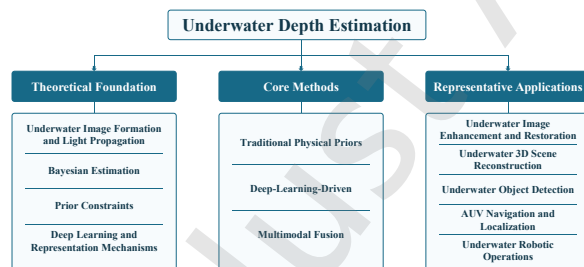


Fig. 1 Organizational framework of our survey.

2 Theoretical Foundation of Underwater Depth Estimation

Underwater depth estimation is essentially an inverse problem that recovers three-dimensional scene structures from degraded two-dimensional observations. Unlike depth estimation in air media, underwater observations are jointly affected by scene structures and camera geometry as well as light absorption, forward scattering, backscattering, wavelength-dependent attenuation, nonuniform illumination, and medium refraction. Consequently, there is generally no simple and stable one-to-one correspondence between scene depth and the

information conveyed in underwater images, such as color, intensity, texture, edges, and disparity.

To elucidate the intrinsic relationship between underwater observations and scene depth, this section introduces the theoretical foundations of underwater depth estimation from four aspects. First, the theory of underwater image formation and light propagation is presented, with emphasis on the effects of light attenuation, scattering, and medium refraction on image observations and depth recovery. Second, Bayesian estimation theory is introduced to explain how observations and prior constraints can be probabilistically integrated to estimate underwater scene depth. Third, commonly used priors involving color attenuation, blurriness, and spatial continuity are summarized. Finally, the role of deep learning in underwater depth estimation is outlined from the perspectives of feature representation and supervisory signals, laying the foundation for the subsequent categorical discussion of traditional physical-based priors, deep-learning-driven approaches, and multimodal fusion methods.

2.1 Underwater Image Formation and Light Propagation

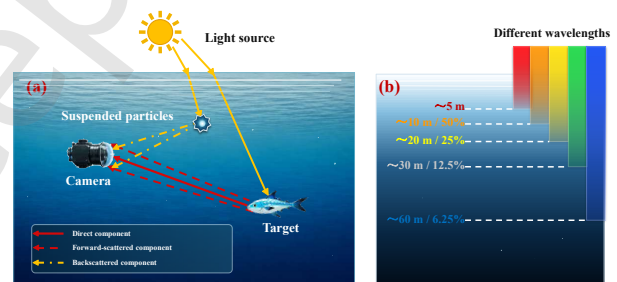


Fig. 2 Schematic diagram of underwater image formation and wavelength-dependent light attenuation.

The schematic diagram of underwater image formation and wavelength-dependent light attenuation is illustrated in Fig. 2. Fig. 2(a) depicts the light entering the water, being reflected by a target, and propagating toward the camera. During propagation, the absorption of light by the water reduces the target-reflected signal that reaches the camera; the forward scattering alters the propagation direction of target light, spreading energy to neighboring pixels and resulting in blurring of edges and textures; the backscattering forms an additional veiling-light component as ambient light is scattered by water and suspended particles into the camera. As the target-to-camera distance increases, the propagation path of the target-reflected light becomes longer, the target information reaching the camera gradually weakens, and the scattering effects accumulate. Consequently, underwater images typically exhibit color shifts, reduced contrast, blurred edges, and loss of fine details. Fig. 2(b) reflects the differential attenuation of light at different wavelengths by the water^[27]. Long-wavelength red light generally attenuates rapidly in water, whereas the blue and green spectral windows propagate over relatively longer distances, thus underwater images tend to exhibit a blue-green color cast. The actual attenuation order of different wavelengths also depends on water type, suspended

particles, dissolved substances, and illumination conditions, so fixed color relationships cannot be regarded as universal rules applicable to all water bodies. Nevertheless, the differential variations among color channels with propagation distance can still provide useful physical cues for underwater relative-depth estimation^[28, 29]. To further elucidate the formation process of underwater images and its relationship with scene depth, we first introduce the Jaffe-McGlamery underwater image formation model, and then analyze the effects of light attenuation, refraction, and backscattering on underwater depth estimation.

The Jaffe-McGlamery model, grounded in the principles of underwater light propagation, decomposes the light received by the camera into three components: the direct component, the forward-scattering component, and the backscattering component. Among them, the direct component represents the portion of the target-reflected light that reaches the camera after attenuation through the water body, forward scattering primarily blurs the target's contours and high-frequency details, whereas backscattering forms a veiling background component that reduces target-to-background contrast. To emphasize the relationship between image observations and scene depth, some studies neglect explicit modeling of forward scattering and adopt the following simplified underwater imaging expression^[8]:

$$I_c(x) = J_c(x)e^{-\beta_c d(x)} + A_c \left(1 - e^{-\beta_c d(x)}\right), \quad (1)$$

where $I_c(x)$ denotes the observed intensity of pixel x in color channel c , $J_c(x)$ is the ideal scene radiance unaffected by water degradation, $d(x)$ is the scene depth corresponding to the pixel, β_c is the effective attenuation coefficient in channel c , and A_c is the backscattered-light intensity at long range. The first term is the attenuated direct component, in which $e^{-\beta_c d(x)}$ is the medium transmission in channel c and describes the absorption and scattering loss over the propagation distance $d(x)$. The second term is the backscattered component, which increases with distance and asymptotically approaches A_c . Since attenuation coefficients differ across color channels, increasing target distance leads to changes in terms of color ratios, contrast reduction, and loss of structural details. This behavior is consistent with the exponential attenuation of light intensity described by the Beer-Lambert law^[29]. Through the direct-signal attenuation and backscattering terms, this model quantifies how the recorded signal changes with depth and establishes a physical relationship among underwater image intensity, color degradation, and scene depth.

Note that Eq. (1) is a simplified representation of real underwater image formation. This expression generally assumes that the water body is approximately homogeneous, that the background light remains spatially consistent, and that the same effective coefficient is used to describe both the attenuation of direct scattering and the variation of backscattering. Akkaynak *et al.*^[30] pointed out that direct scattering attenuation and backscattering have different formation mechanisms and should be described using different parameters; in scenes with artificial illumination, non-uniform water bodies, or locally highlighted regions, background scattering may also vary with spatial position. Furthermore, forward scattering is not represented separately in Eq. (1), and its effects are usually

subsumed into image blur or effective scene radiance. Color and intensity therefore provide useful distance cues, but are simultaneously affected by object reflectance, water type, and illumination distribution, preventing them from determining accurate scene depth in isolation.

In addition to the radiometric changes caused by absorption and scattering, medium refraction also alters the geometric relationships in underwater imaging. Underwater cameras typically observe the external scene through flat or curved waterproof housings. Light passes sequentially through the water, the waterproof glass, and the air inside the housing, and is refracted at the interfaces between media with different refractive indices^[6, 7]. According to Snell's law, the angles of incidence and refraction in the two media satisfy:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2, \quad (2)$$

where n_1 and n_2 are the refractive indices of the two media, and θ_1 and θ_2 are the incidence and refraction angles relative to the interface normal. The light refraction changes the effective projection center, field of view, and focal-length relationship of the underwater camera, thus a pinhole model calibrated in air cannot accurately represent the underwater ray path^[6, 7]. For an ideal pinhole stereo model, the scene depth is calculated from the focal length f , stereo baseline B , and disparity δ as follows:

$$Z = \frac{fB}{\delta}, \quad (3)$$

where Z denotes the scene depth, f is the focal length, B is the stereo baseline, and δ denotes the disparity. However, in underwater environments, if the camera parameters and epipolar geometry calibrated in air are directly used for disparity matching and triangulation, the refraction of the medium will cause systematic depth bias. Stereo and multi-view underwater depth estimation therefore usually requires underwater camera recalibration or an explicit refractive model that corrects the relationship among image pixel coordinates, propagation rays, and scene points. The light propagation model primarily explains how image appearance varies with distance, while the refractive imaging geometry determines how image coordinates are mapped to physical space. Together, they constitute the physical basis for underwater relative depth and metric depth recovery^[6].

In summary, the underwater light-propagation models reveal the relationships among attenuation, scattering, wavelength-dependent differences, and scene depth, providing color and degradation cues related to propagation distance for monocular images. The refractive imaging geometry describes the mapping relationships among underwater pixel observations, ray directions, and 3D scene points, forming the basis for stereo, multi-view, and metric depth recovery. Underwater depth estimation must consider both radiometric and geometric changes to avoid mistaking water-induced degradation for scene structure or propagating refractive errors into the final depth results.

2.2 Bayesian Estimation

Underwater depth estimation recovers dense scene depth from incomplete observations such as color, intensity, texture, disparity, or sparse range measurements, and is therefore a typical inverse problem. Similar underwater image appearances may

correspond to different scene depths, while objects at the same depth may appear markedly different due to differences in object reflectance, water type, and illumination conditions. To jointly incorporate image observations, image-formation models, and existing constraints, Bayesian estimation can be used to perform probabilistic inference on scene depth, whose basic form is:

$$p(d | z) \propto p(z | d) p(d), \quad (4)$$

where d is the scene depth to be estimated and z represents the corresponding observation. The prior distribution $p(d)$ describes the depth constraints, the likelihood $p(z | d)$ expresses the probability of the observations given a candidate depth, and the posterior $p(d | z)$ represents the inferred depth distribution after combining the observation and prior.

In underwater depth estimation, likelihood functions can be established for different observations according to different types of observation models. For methods based on underwater physical priors, the likelihood measures whether a candidate depth can explain the color attenuation, transmission, and backscattering observed in the current image through an underwater imaging model. For stereo or multi-frame methods, disparity consistency, cross-view reprojection error, or photometric reconstruction error can measure the degree of matching between depth and geometric observations. For visual-acoustic, visual-laser, or visual-SLAM fusion approaches, corresponding probabilistic models can be established from the errors between predicted depth and external range or sparse geometric observations.

The prior distribution is used to narrow the space of possible depth solutions. Depths on the surface of the same object are usually relatively continuous, while significant depth changes are allowed at object boundaries. Underwater color attenuation, image blur, and backscattering follow certain regularities with respect to propagation distance, while sonar, laser, and sparse map points can further provide the actual scale range of depths. By representing this information as prior distributions, the depth ambiguity caused by relying solely on a single observation can be reduced. The posterior distribution gives the probabilities of different candidate depths under the current observations. To obtain a determinate depth result, maximum a posteriori (MAP) estimation is commonly employed, selecting the depth with the highest posterior probability:

$$d^{\text{MAP}} = \arg \max_d p(d | z) = \arg \max_d p(z | d) p(d). \quad (5)$$

Taking the negative logarithm of Eq. (5), this MAP estimation can be reformulated as an energy minimization problem consisting of a data likelihood term and a prior regularization term. The former requires the predicted depth to explain the current image or multimodal observations, while the latter constrains the prediction results to satisfy reasonable spatial structures, physical laws, or scale ranges. Supervised losses, photometric reconstruction losses, smoothness losses, and physical-consistency losses in deep learning-based models can all be interpreted as specific realizations of the likelihood term or the prior term within this framework.

Beyond a single depth prediction, the Bayesian framework can also characterize the uncertainty of the prediction.

When the posterior distribution is relatively concentrated around a certain depth, it indicates that the current observations impose strong constraints on that region; when the posterior distribution is relatively dispersed, it suggests that color, texture, geometric, or external distance information is insufficient, and the prediction is therefore subject to high uncertainty. This uncertainty is particularly important for underwater robot avoidance, obstacle navigation, and precise manipulation, because the system must not only output depth but also assess whether the depth estimate is reliable. The relationship among Bayesian observation models, priors, and posterior inference has been demonstrated in underwater single-photon light detection and ranging (LiDAR) depth recovery. For example, by incorporating sparse photon observations and depth priors into a Bayesian model^[31] to improve depth recovery under signal attenuation and noise interference, and subsequently by adopting a hierarchical Bayesian model to fuse multi-scale information^[32], thereby further enhancing the robustness of depth reconstruction under complex degradation conditions. In summary, Bayesian estimation provides a unified probabilistic interpretation of underwater depth estimation: the likelihood measures how well a candidate depth explains the current observations, the prior constrains the plausible depth range and structure, and the posterior combines the two to determine both the final depth and its uncertainty.

2.3 Prior Constraints

The term prior is used broadly in the underwater depth estimation. Strictly speaking, spatial continuity, depth range, and scale distributions can be represented directly by the depth prior $p(d)$ in Bayes' rule, whereas regularities such as color attenuation, dark channel, blurriness, and backscattering describe empirical relationships between image observations and propagation distance, and can be used to construct observation models, generate initial depth, or serve as auxiliary network inputs. Commonly used priors in underwater depth estimation include the color attenuation prior, the blurriness prior, and the spatial continuity prior, whose basic mechanisms of action are illustrated in Fig. 3.

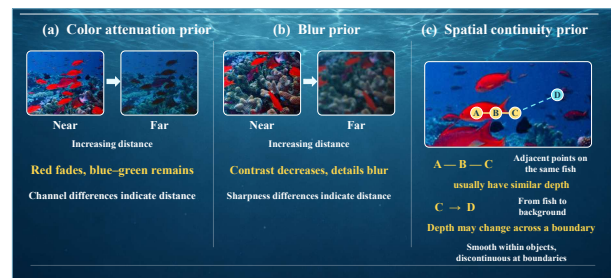


Fig. 3 Three types of priors commonly used in underwater depth estimation.

(1) *Color-attenuation prior.* This type of prior arises from the wavelength-dependent absorption and scattering of light in water. As shown in Fig. 3(a), as the target distance increases, the intensity and relative proportions of different color channels change differently, therefore, the discrepancies

among the red, green, and blue channels can serve as cues for judging the relative distance of the target [28]. This prior can extract propagation-distance information from a single image, but its reliability depends on water types, illumination conditions, and intrinsic object colors. When the target itself has a pronounced blue, green, or red appearance, the object color may be confounded with water attenuation, consequently, color attenuation is generally more suitable for providing relative depth or rough distance estimates, while it is difficult to independently recover accurate metric depth. (2) *Blurriness prior*. This prior considers the relationship between image degradation and target distance. As shown in Fig. 3(b), the farther the target distance, the longer the propagation path of light in water, and the more pronounced the cumulative effects of absorption and scattering. Consequently, image contrast decreases, and details such as edges and textures gradually weaken [33, 9]. Since image blurriness can also be affected by factors such as camera motion and imaging defocus, this prior usually needs to be used in combination with other depth cues. (3) *Spatial continuity prior*. Color attenuation and blurriness provide the cues for ordinal range, while scene structure can further constrain the depth map. In general, depth variations within the same object surface or a locally continuous region tend to be relatively smooth, whereas significant depth changes may occur at the boundaries between different objects. As shown in Fig. 3(c), the spatial-continuity prior can be used to reduce locally unreasonable depth fluctuations while preserving depth discontinuities at object boundaries.

Additionally, for stereo, multi-view, and video-based depth estimation, cross-view geometric consistency can also be introduced. A correct depth map should enable projecting pixels from one view to their corresponding positions in another view by incorporating the camera's intrinsic and extrinsic parameters, and the predicted depth should also agree with stereo disparity, inter-frame camera motion, and sparse 3D map points. Compared with color and blur priors, geometric consistency relies less on object color and provides a more direct basis for metric scale recovery, but its reliability depends on camera calibration, pose estimation, corresponding point matching, and the accuracy of the underwater refraction model.

In summary, prior constraints alleviate the problem of recovering scene depth from a single observation by incorporating underwater light-propagation laws, image degradation characteristics, and scene spatial structures.

Color attenuation and blurriness priors mainly provide relative depth cues related to underwater propagation distance, spatial continuity regularizes local structures of the depth map, and cross-view geometry or external distance information can supplement scale and 3D positional constraints. Since any single prior may fail in a specific scene, practical methods usually require joint modeling of multiple priors and dynamic adjustment of different constraints according to water body, illumination, texture, and observation quality. Meanwhile, the applicability of priors is also affected by environmental conditions and imaging devices. Therefore, underwater depth estimation requires not only exploiting priors, but also assessing the reliability of each prior in the current image to avoid systematic biases in depth results caused by erroneous priors.

2.4 Deep Learning and Representation Mechanism

Deep learning-based methods establish nonlinear mappings from underwater images to pixel-wise depth through multi-layer networks, and progressively learn representations from local details to global structures. Three types of models, namely CNNs, Transformers and Mamba, are presented on their roles in terms of local feature extraction, long-range dependency modeling, and computational efficiency for underwater depth estimation.

CNN. CNNs extract features from local image neighborhoods by convolution and gradually enlarge their receptive fields through stacked layers, forming hierarchical representations from edges and textures to object contours and structure. Local CNN features can help identify underwater object boundaries and fine structural variations. For example, a convolutional depth network [19] is trained using reprojection between adjacent frames without underwater ground-truth depth. UWDepth [33] uses a convolutional network to exploit inter-frame geometry via an adaptive cost volume and depth-consistency constraints. These demonstrate that CNNs are effective in extracting local edges, textures and contours for underwater depth estimation, but have limited modeling on long-range dependencies.

Transformer. Transformer uses self-attention to compute the relationships among features at different positions and establishes long-range dependencies within a single layer. Unlike CNNs that enlarge the receptive field gradually, Transformer can relate distant regions and help infer ordinal relationships from overall scene context. Transformers are usually combined with CNNs for underwater depth estimation. For example, UMono [34] adopts parallel CNN and Transformer encoders to extract local and global features, and combines these features with underwater medium-transmission information for depth prediction. And CTFDepth [9] contains a CNN depth-prediction branch and a Transformer image-reconstruction branch, and the contextual information from the reconstruction branch is fused into depth features. These methods use CNNs for local details and adopt Transformers for long-range context, thereby combining local structures with global scene information.

Mamba. Transformer models long-range dependencies, but its self-attention has quadratic computational complexity with respect to feature-sequence length. Mamba [35] uses a selective state-space mechanism to propagate information with linear complexity, providing an alternative for large-scale feature modeling. Mamba converts two-dimensional features into sequences according to scan orders, but fixed orders cannot effectively represent spatial relationships among underwater objects. Tree-Mamba [18] constructs tree-structured connections according to feature similarity and aggregates information, which uses broad scene context with controlling computation, balancing global modeling and efficiency in underwater depth estimation. In practice, CNNs, Transformers, and Mamba can work in concert, jointly exploiting local details and global information of underwater images to accomplish scene depth prediction.

3 Core Methods for Underwater Depth Estimation

This section presents the core methods for underwater depth estimation, including three categories: traditional physical prior-based methods, deep-learning-driven methods, and multimodal fusion methods. Table 1 summarizes these core methods for underwater depth estimation, among them, traditional physics priors and deep learning mainly differ in the depth inference mechanism, whereas multimodal fusion emphasizes the introduction of external distance or geometric cues. Besides, these three types of techniques have shown a clear trend of cross-integration in recent years.

3.1 Traditional Physical-Prior Methods

In underwater depth estimation, as the propagation distance increases, the absorption and scattering of light by the water body cause changes such as color attenuation, contrast degradation, image blurring, and backscattering. Traditional physical priors treat these changes as depth cues and estimate depth by establishing the relationship between image features and propagation distance. Existing traditional physical-prior methods can be broadly categorized into four categories: dark channel priors, light attenuation and color channel priors, blurriness priors, and backscattering priors. The specific subdivisions of traditional physical prior methods are shown in Table 2.

Dark-Channel Priors: Dark-channel priors aim to estimate transmission from local dark-pixel variations and infer scene depth from the relationship between transmission and propagation distance. As target-reflected light attenuates and backscattering becomes heavy, local dark pixels in the observed underwater image generally become brighter. Since red light attenuates rapidly and red-channel intensities are uniformly low, the Underwater Dark Channel Prior (UDCP) estimates transmission using local dark pixels in green and blue channels^[14, 36]. The transmission $\hat{t}(x)$ estimated by UDCP can be expressed by

$$\hat{t}(x) = 1 - \min_{y \in \Omega(x)} \left(\min_{c \in \{G, B\}} \frac{I^c(y)}{A^c} \right), \quad (6)$$

where $\Omega(x)$ is a local window centered at image pixel x , y denotes a pixel within $\Omega(x)$, G and B denote green and blue channels respectively, $I^c(y)$ is the observed intensity at y in channel c , and A^c is the background-light intensity^[36]. Transmission is the fraction of target-reflected light that reaches the camera after passing through water. Lower transmission usually indicates a longer propagation path and a more distant target, and their conversion relationship^[14] can be expressed as

$$d = -\frac{\log(t)}{\beta_c}, \quad (7)$$

where d denotes the scene depth, t is the estimated transmission, and β_c is the attenuation coefficient of channel c . Dark channel-based methods therefore estimate transmission from local dark pixels and then calculate scene depth from the estimated transmission.

Vaz *et al.*^[14] first classified water types and then selected a learning-based or model-driven method to recover underwater depth from transmission and attenuation coefficients. Guo

et al.^[36] formed a light-attenuation input from grayscale intensity, maximum green/blue response, UDCP transmission, and fused its predicted depth with that extracted from underwater red-green-blue (RGB) images. Zhou *et al.*^[37] estimated underwater depth using a channel-intensity prior, and handled backscattering through adaptive dark-pixel processing. In these methods, dark channel priors connect local dark pixels, transmission, and scene depth. They require no ground-truth depth and provide intuitive model-based inference. But their results depend on accurate local dark pixels, background light, and attenuation coefficients, and can be biased under complex lighting or unusual color conditions. Thus they are useful for coarse depth inference or as physical attenuation cues combined with other knowledge.

Light-Attenuation and Color-Channel Priors: Light-attenuation and color-channel priors can exploit wavelength-dependent absorption to extract underwater depth cues from distance-dependent color changes. Tamaki *et al.*^[38] used a physics-based depth-calculation method based on color-channel values, where the red-channel component and the larger value of green and blue channels are used to calculate underwater image depth as follows:

$$d(x; R, M) = \mu_0 + \mu_1 R(x) + \mu_2 M(x), \quad (8)$$

where x denotes the pixel location, $d(x; R, M)$ denotes the corresponding depth, $R(x)$ denotes the red-channel component, $M(x)$ denotes the larger value of green and blue channels, μ_0 , μ_1 , and μ_2 are preset parameters. This method assumed that the intensity difference between the red channel and the green/blue channels varies with target range, obtaining depth by a simple calculation without training a network, but its parameters were determined in advance and applicability was affected by water and illumination conditions. To improve the robustness to brightness and color casts, Zhou *et al.*^[37] proposed the channel intensity prior (CIP), and produced one relative depth map from red-channel attenuation and another from inter-channel differences, then weighted them according to global brightness:

$$d_{CIP} = \alpha d_m + (1 - \alpha) d_r, \quad (9)$$

where d_{CIP} denotes the depth estimated by the channel intensity prior, d_r denotes the depth from red-channel attenuation, d_m denotes the depth from inter-channel differences, and α is the weight determined by image brightness. Subsequently, Zhou *et al.*^[37] fused the channel-prior depth with that predicted by an unsupervised model according to the degree of color cast, reducing failures caused by any single channel cue. Additionally, channel-attenuation information can also be used as a network input. Guo *et al.*^[36] constructed an IMT input space containing grayscale intensity, green/blue channels, and transmission, and predicted depths from RGB and IMT inputs. RGB provides color and texture features, while IMT represents the relationship between attenuation and propagation distance. Then reliability-aware fusion was used to supplement weak cues in distant blurred regions. Xiao and Qu^[39] used inter-channel attenuation relationships to constrain self-supervised depth estimation. Ratios among red, green, and blue channel attenuation coefficients were bounded using inherent optical properties, while image restoration, re-degradation, and inter-frame projection were combined so that the predicted depth satisfied both underwater optics and

Table 1 Core methods for underwater depth estimation

Category	Manner	Core	Advantage	Limitation
Traditional physical priors	Dark channel, light attenuation and color channels, blurriness, backscattering	Extract depth cues from underwater propagation, channel attenuation, blur, and scattering	High interpretability and less dependence on labeled data	Sensitive to water quality, illumination, object color, and prior mismatch
Deep-learning-driven	CNN, Transformer, Mamba	Learn nonlinear mappings from underwater images to depth maps	Strong representation ability and high accuracy in complex underwater scenes	Dependent on data and computation, with limited generalization
Multimodal fusion	Vision–acoustic, vision–laser, vision–SLAM	Introduce external range measurements, sparse depth, or temporal geometry	Stable scale and explicit geometric constraints	Complex and costly hardware, with difficult calibration

Table 2 Underwater depth estimation methods based on traditional physical priors

Prior	Literature	Cue	Core	Advantage	Limitation
Dark channel priors	Vaz <i>et al.</i> [14]	Local dark pixels, transmission, attenuation coefficient	Select the model by water type and compute depth from transmission and attenuation	Physically meaningful inversion	Depends on water types, limited accuracy
	Guo <i>et al.</i> [36]	RGB channels, grayscale, UDCP transmission	Predict from RGB and IMT fusion by confidence	Combine learned appearance with physical attenuation	Unreliability by severe degradations or nonuniform illumination
	Zhou <i>et al.</i> [37]	Channel intensity, dark pixels, learned depth	Combine channel priors with Monodepth2 and remove backscattering	Improved depth estimation under color cast and complex lighting conditions	Absolute scale depends on an empirical range, relatively high computation
Light-attenuation and color-channel priors	Tamaki <i>et al.</i> [38]	Red channel, maximum of green/blue	Compute depth from channel intensities	Training-free single image estimation	Preset parameters, sensitive to environments
	Zhou <i>et al.</i> [37]	Red channel, channel difference, image brightness	Weigh two channel cues for depth maps	Reduces failure of single channel cue	Relative depth acquisition, relies on empirical parameters
	Guo <i>et al.</i> [36]	Grayscale, RGB channels, transmission	Fuse RGB and IMT predictions	Supplements weak cues in distant blurred regions	Low contrast and uneven illumination
	Xiao and Qu [39]	Attenuation coefficients inter-frame geometry	Take optical and reprojection constraints for self-supervised training	Without ground truth	Require video with relatively complex model
Blurriness priors	Peng <i>et al.</i> [12]	Multi-scale blurriness	Generate, fill, and smooth coarse depth maps	Better relative depth	Errors in smooth and weak texture regions
	Peng <i>et al.</i> [11]	Blurriness, red channel, channel differences	Combine three depth estimates using background and red information	Fusion of sharpness and color attenuation	Depend on background light, threshold, and scale
Backscatter priors	Varghese <i>et al.</i> [20]	Direct and backscatter transmission	Infer depth from transmission maps and adjacent frames	Backscatter participates in inversion	Requires video training, errors in turbid or specular regions
	González-Sabbagh <i>et al.</i> [40]	Backscattering, formation model	Jointly predict restoration, scattering and depth by reconstruction error	Consistency with image formation	Rely on synthetic RGB-D data
	Du <i>et al.</i> [41]	Relative depth, scattering, optical parameters	Re-degrade predicted quantities and optimize via reconstruction	Backscatter constraints and depth correction	Not direct depth prediction, Relatively high computation

temporal geometry. Overall, light-attenuation and channel priors relate underwater color variations with object range, and can be used to compute relative depth or provide learned models with useful information about underwater propagation.

Blurriness Priors: In addition to image intensity and color, image sharpness varies with underwater range. Peng and Cosman [11] computed an initial blurriness map by comparing a grayscale image with Gaussian-filtered versions at multiple scales, and its expression is as follows:

$$P_{\text{init}}(x) = \frac{1}{n} \sum_{i=1}^n |I_g(x) - G_{r_i, r_i}(x)|, \quad (10)$$

where $P_{\text{init}}(x)$ denotes the initial blurriness value at pixel x , $I_g(x)$ denotes the grayscale intensity at pixel x , $G_{r_i, r_i}(x)$ denotes the pixel value after Gaussian filtering at the i th scale, $r_i = 2i/n + 1$ denotes the Gaussian-filter scale, and n denotes the number of scales. Eq. (10) extracts local high-frequency information of underwater images by calculating the difference between the raw grayscale image and results

obtained by Gaussian filters at different scales. A smaller difference indicates weaker edges and textures and thus greater blurriness. Because edges, textures, and details of underwater targets gradually weaken as propagation distance increases, image blurriness can be used to infer relative scene range. Peng *et al.* [12] applied this cue to single-image underwater depth estimation. Pixel-wise blurriness was computed from multi-scale Gaussian filtering, maximum filtering was used to generate a coarse depth map, and smooth regions lacking strong edges were filled before final smoothing. Blurriness alone is ambiguous: a nearby smooth object may appear blurred because it lacks texture and may therefore be classified as distant. Peng and Cosman [11] supplemented blurriness with channel changes caused by absorption, generating three depth estimates from red-channel intensity, the difference between red and green/blue channels, and blurriness. The blurriness characterized the weakening of edges and details with propagation distance, and range was inferred mainly from blurriness when red content was weak. Therefore, blurriness priors can exploit the weakening of underwater image edges and details with increasing distance to infer range and provide depth cues distinct from color channels, but are prone to errors in smooth or weakly textured regions, generally combined with light absorption and channel-attenuation to improve the reliability of depth estimation.

Backscatter Priors: Backscatter priors focus on ambient light scattered by suspended particles into the camera, forming an additive background component. Varghese *et al.* [20] constructed a direct-light transmission map and a backscatter transmission map, and their relationships with scene depth are shown as follows:

$$\begin{aligned} T_c^d(x) &= e^{-\beta_c^d D(x)}, \\ T_c^b(x) &= e^{-\beta_c^b D(x)}, \\ D(x) &= -\frac{\ln T_c^d(x)}{\beta_c^d} = -\frac{\ln T_c^b(x)}{\beta_c^b}, \end{aligned} \quad (11)$$

where $D(x)$ denotes the scene depth at pixel x , $T_c^d(x)$ and $T_c^b(x)$ denote the direct-signal transmission and backscatter transmission in channel c respectively, β_c^d and β_c^b are corresponding attenuation coefficients. As scene depth increases, transmissions decrease, the target-reflected component weakens, while the background-light by backscatter becomes stronger. Therefore, scene depth can be calculated from the relationship between transmission and propagation distance.

Backscatter information is used in two principal ways: one calculates the depth from direct-signal and backscatter transmission, and the other incorporates backscattering into underwater image reconstruction and constrains predicted depth through the reconstructed result. Representative works are shown in Table 2. In the first category, USe-ReDI-Net [20] decomposed an underwater image into scene radiance, direct transmission, backscatter transmission, background light, and further used adjacent-frame geometry to refine the depth during training. Unlike methods that calculate depth directly from transmission, Scene-cGAN [40] predicted the restored image, backscatter, and scene depth separately. These quantities are inserted into an underwater image-formation model, and the difference between generated and input images constrains

network training and enforces predicted depth consistent with image formation model, where backscatter contributes to depth constraint by image reconstruction. Du *et al.* [41] estimated scene depth, backscatter intensity, attenuation and scattering coefficients of light jointly. This method considered spatially nonuniform backscatter caused by artificial illumination, and backscatter served to constrain and correct depth estimation since initial one came from a pretrained depth network. Backscatter priors participate in underwater image reconstruction to constrain predicted depth. Since backscatter and target-reflected light coexist in underwater observations and are difficult to separate accurately, these methods commonly combine transmission, inter-frame geometry, or learned models for underwater depth estimation.

Overall, dark channel priors mainly use transmission to support depth inversion, light-attenuation and color-channel priors exploit distance-dependent changes among color channels, blurriness priors focus on phenomena such as weakened edges, reduced textures, and blurred details in distant regions, and backscatter priors utilize the characteristic that scattered light intensifies with increasing distance. Although these methods use different cues, they all extract distance-related information from underwater image degradations to estimate depths.

3.2 Deep-Learning-Driven Methods

The aforementioned physical prior methods infer scene depth based on underwater imaging principles such as color attenuation, image blurring, and backscattering, and thus possess strong interpretability. However, they have limited adaptability to complex degradations, which has made deep learning methods the mainstream. Different methods have formed distinct technical routes in terms of the depth information required for training and the approach to network feature extraction. Therefore, deep-learning-driven methods can be categorized by learning paradigm and model type, as summarized in Table 3.

Supervised Learning

Supervised underwater monocular depth-estimation methods are usually trained on paired underwater images and ground-truth depth labels to learn a mapping from an image to its depth map. Through end-to-end learning, these methods extract depth cues associated with color, texture, and degradation features in underwater images, and thus have strong depth-recovery capability.

CNN-Based Supervised Learning: Underwater scattering and blur weaken object boundaries, leading to detail loss of depth maps. To address the above problem, Chen *et al.* [42] proposed UW-LapDepth, which uses a Laplacian pyramid to progressively predict depth information at different scales. This model first recovers overall scene-depth distribution and then gradually supplements object boundaries and local details, thereby reducing boundary blurs in the depth map. After basic scene depth can be recovered, some studies further address image degradation and the reliability of estimated depth. Yang and Cho [43] presented TRIDENT, which predicts scene depth and estimates depth uncertainty simultaneously. The enhanced underwater image provides clearer edges and structures for

Table 3 Deep-learning-driven underwater depth estimation methods

Learning	Architecture	Literature	Data	Core	Advantage	Limitation
Supervised learning	CNN	UW-LapDepth ^[42] , TRIDENT ^[43] , UDepth ^[15]	Paired underwater images and depth labels	Extract local texture and edge features with convolution	Strong local-detail modeling and clear structures	High-quality depth labels, limited generalization
	Transformer	Joint-ID ^[16] , UPGformer ^[17] , CD-UDepth ^[36]	Paired underwater images and depth labels	Model global context with self-attention and physical priors	Strong global structure under complex degradations	High computational complexity
	Mamba	Tree-Mamba ^[18] , UDepthDiff ^[44]	Paired underwater images and depth labels, pseudo-depth labels	Model long-range dependencies with state-space and tree-scanning	Long-range modeling and efficiency	Strong dependence on datasets and architecture design
Unsupervised Learning	CNN	UW-Net ^[45] , CIP+ ^[37] , UWdepth ^[33] , USe-ReDI-Net ^[20] , Monodepth2 ^[46]	Unpaired underwater images and terrestrial RGB-D data, underwater image sequences	Build training signals from domain mapping, reprojection or consistency	Relatively good applicability	Sensitive to scattering, color cast and illumination
	CNN-Transformer	CTFDepth ^[9] , WaterMono ^[47]	Underwater image sequences	Combine inter-frame variation with global structures	Local details and global context	Higher training and inference cost

depth prediction, while uncertainty indicates the confidence of depth result in each region. In addition to improving underwater depth estimation accuracy, computational cost is also considered. Yu *et al.*^[15] designed UDepth, which uses lightweight MobileNetV2 as its main network and incorporates underwater light-attenuation information to assist depth prediction. It focuses on retaining effective underwater range cues while controlling model size, making it suitable for underwater devices with limited computing capability. These above methods improve CNN-based estimation through multiscale detail restoration, multitask learning, or lightweight design, but rely on local information and are less effective at relating large-scale scene regions.

Transformer-Based Supervised Learning: Since distant underwater regions usually exhibit similar colors and weak textures, CNNs relying on local features lead to confusion between the ranges of different objects. Transformers link separated image regions and infer spatial relationships from overall scene structures, complementing the limitation of CNNs in global information modeling. Yang *et al.*^[16] proposed Joint-ID, which introduces Transformers into a joint underwater image enhancement and depth estimation framework in a shared hierarchical feature space. This network shows more stable depth representations under low contrast, color casts, and underwater degradation, demonstrating the value of Transformers in multi-task underwater visual perception. To address this problem of learning global structures alone that causes underwater color changes mistaken for intrinsic object features, Wang *et al.*^[17] introduced UPGformer that integrates a Transformer with underwater image formation information. The Transformer links global structures of different regions, while the transmission supplies range cues related to light attenuation, improving underwater depth estimation stability in distant regions/areas with spatially nonuniform degradations. Liu *et al.*^[48] addressed scale variations caused by underwater refraction, low contrast, and nonuniform illumination by introducing an intrinsic scale-invariance mechanism into a Vision Transformer encoder. Local planar guidance and skip-expanded context fusion in the decoder improve predictions for underwater objects of different scales, boundaries, and local regions. To further account for spatial differences in the reliability of color and light-attenuation, Guo *et al.*^[36]

presented CD-UDepth to predict depth from both RGB images and IMT information composed of grayscale intensity, green/blue channels and transmission. And a lightweight Vision Transformer supplements large-scale underwater scene structures, allowing the model to select reliable depth cues according to image degradations. In summary, supervised Transformer-based approaches incorporate image enhancement, transmission, and multi-source feature fusion into global modeling, and their advantage lies in jointly reasoning about overall scene structures. However, these Transformers have quadratic computational complexity, and remain limited by the scale of training data and the quality of depth labels.

Mamba-Based Supervised Learning: Compared with Transformers that rely on self-attention for global modeling, Mamba has an advantage in computational complexity. When Mamba is used for image processing, two-dimensional features must be organized into a sequence along a specified path. Common horizontal, vertical, and diagonal scan paths are manually predefined, which cannot adapt to object shapes or scene structure in underwater images. When contiguous regions of the same object are separated by the scan path, the model fails to preserve structural integrity, and results in blurred depth boundaries or erroneous local depths. To address this problem, Zhuang *et al.*^[18] proposed Tree-Mamba, which dynamically determines feature-propagation paths from underwater image content. CNNs first extract multi-scale local features, after which tree connections are constructed according to the feature similarity of adjacent regions so that the information can propagate along object structure. Compared with fixed-direction scanning, this method preserves better object contours and regional continuity, and reduces boundary blurs and missing local details in depth maps. To alleviate the shortage of underwater depth data, Zhuang *et al.*^[18] constructed BlueDepth, containing 38,162 underwater image-pseudo-depth pairs. The depth labels were generated by multiple monocular depth-estimation models and subjected to quality screening. Unlike Tree-Mamba, UDepthDiff^[44] employs Mamba to improve the physical information that assists depth estimation. Mamba processes intensity variations and structural information in underwater images, and fuses extracted object regions with inverse-transmission information. The fused result serves as an auxiliary condition that guides the

diffusion model in progressively generating a depth map, improving depth discrimination among foreground, background, and object boundaries. Overall, Mamba-based methods offer a new modeling approach for underwater depth estimation. Tree-Mamba combines local feature-extraction capability of CNNs with Mamba's long-range modeling, and uses the tree-aware scanning to preserve local continuity and spatial topology. UDepthDiff further applies Mamba scanning to physical-prior generation in frequency domain, using diffusion-based depth estimation to improve prediction stability under complex degradations.

Although all above supervised learning methods can achieve strong depth estimation performance, they depend on large quantities of high-quality depth labels. Meanwhile, real underwater depth is difficultly acquired, and limited scene coverage restricts model generalization.

Unsupervised Learning

Different from the above supervised learning, unsupervised learning methods do not depend on manually annotated depth. Instead, they learn depth cues from image structures, reconstruction constraints, cross-domain translation, and consistency constraints. When real depth labels are unavailable, these underwater depth estimation methods can exploit color attenuation, illumination changes and structural relationships, to improve model applicability in complex aquatic environments.

CNN-Based Unsupervised Learning: CNN-based unsupervised learning uses convolutional networks to learn the relationship between underwater images and scene depth, reducing dependence on paired underwater red-green-blue-depth (RGB-D) data. Gupta *et al.* [45] proposed UW-Net, which uses dense-block convolutional autoencoders and cycle-consistent learning to establish a mapping between underwater images and unpaired terrestrial RGB-D data. This method does not require underwater depth labels, but is affected by the training-data distribution. Monodepth2 [46] was established on monocular video training without ground-truth depth, which can predict scene depth and camera motion, warp adjacent frames into a common viewpoint, and train from photometric reconstruction error. When such methods are applied directly to underwater images, color attenuation and brightness variations cause reconstruction errors to fail in reflecting depth errors accurately. Zhou *et al.* [37] presented CIP+, which combines depth cues obtained from channel-intensity variations with Monodepth2 predictions and uses underwater light-attenuation information to supplement inter-frame geometry. Building on this, Wang *et al.* [33] introduced UWdepth, which uses multiple consecutive images to estimate depth jointly, and compares changes in the same region across frames and corrects depth, thereby strengthening inter-frame geometry. Besides, USe-ReDI-Net [20] incorporated underwater image restoration into unsupervised training. This model first predicts scene depth, restores a clearer image using underwater image-formation model, then checks whether the restoration is consistent across adjacent frames. Overall, these CNN-based methods reduce dependence on underwater RGB-D pairs via unpaired cross-domain mapping or temporal geometry, and some of them incorporate attenuation cues or image formation models to validate and constrain depth.

Hybrid CNN-Transformer Unsupervised Learning: Un-

supervised methods introduce Transformers to supplement global structural information after CNNs extract local depth features. Representative methods combine CNNs with Transformers rather than using a pure Transformer depth network. Cheng *et al.* [9] proposed CTFDepth for underwater multi-frame monocular depth estimation. The CNN branch extracts object edges and positional changes from consecutive images to predict scene depth, while the Transformer branch masks image parts and reconstructs them from surrounding content. After the two branches are fused, local edges and global structures jointly contribute to improving depth prediction in complex degraded regions, but continuous video may contain moving fish and changes in camera pose. To address the limitation, Ding *et al.* [47] presented WaterMono, which combines CNN with Transformer to learn scene depth from consecutive underwater images. The method first compares adjacent frames to estimate relative object range, and then identifies moving regions to reduce their influence on training. By including underwater images captured from different orientations, this method maintains stable depth estimation when the camera is tilted or inverted. These methods show that alone expanding feature context is insufficient, and self-supervised training must account for underwater appearance changes and actual imaging process.

For deep learning-based underwater depth estimation, supervised methods can be accurate when sufficient data are available but are limited by annotation. Unsupervised methods have lower accuracy but greater practical value in real underwater scenes. Combining CNNs, Transformers and Mamba with physical priors has become an important direction for improving the performance of underwater depth estimation.

3.3 Multimodal Fusion Methods

Different from preceding depth estimation methods based on underwater images, multimodal fusion methods introduce information from other sources, where acoustic and laser sensors provide physical ranges, and SLAM uses consecutive images to establish spatial relationships in the scene. Multimodal fusion jointly uses these sources to complement one another for depth inference. When one source is biased, the others can provide additional evidence and correct the estimation, thereby reducing dependence on any single cue and improving stability in complex underwater environments. Existing methods for multimodal fusion can be categorized into three types: main visual-acoustic fusion, visual-laser fusion, and visual-SLAM fusion, with representative studies summarized in Table 4.

Visual-Acoustic Fusion: Acoustic sensors are relatively insensitive to optical degradations, which provide stable ranges in turbid, dim, or distant scenes, and their combination with monocular vision supplies external scale constraints to mitigate the ambiguity. Early research employs acoustic range measurements to correct the scale of monocular vision. Roznere *et al.* [5] associated ranges from a single-beam echo sounder with visual features to correct scene scale recovered by the visual system. Mantani *et al.* [23] extended acoustic information from global scale correction to local depth-map correction, where a monocular model generates a depth map after which sonar target ranges correct corresponding regions. Liu *et al.* [24] incorporated single-beam echo-sounder measurements into self-supervised training, and converted individual ranges

Table 4 Categories of multimodal fusion methods

Category	Method	Literature	Characteristic
Visual-acoustic fusion	Global scale correction	Roznere <i>et al.</i> [5]	Associate single-beam ranges with image features to correct global scale recovered by monocular vision
	Local depth correction	Mantani <i>et al.</i> [23]	Correct corresponding regions of depth map using sonar-measured target ranges
	Training constraint	Liu <i>et al.</i> [24]	Convert acoustic ranges into regional depth prompts for self-supervised depth-model training
	Cross-modal alignment and feature fusion	Lv <i>et al.</i> [25]	Align and fuse visual texture features with sonar-scale features to predict dense metric depth
Visual-laser fusion	Single-frame local depth measurement	Campos <i>et al.</i> [49]	Detect laser patterns on object surfaces and obtain local depth in single image by triangulation
	Continuous depth measurement	Inglis <i>et al.</i> [50], Bodenmann <i>et al.</i> [51]	Acquire laser-line depth profiles to recover metric depth over larger areas
Visual-SLAM fusion	Sparse geometric prior-assisted prediction	Ebner <i>et al.</i> [4]	Fuse sparse SLAM map points with RGB image to predict dense depth
	Sparse scale correction and refinement	Zhang <i>et al.</i> [52]	Use sparse SLAM depth to correct relative-depth scale and generate dense metric depth

into regional depth prompts and combined with consecutive images, reducing absolute-depth errors. Lv *et al.* [25] presented RSFNet that converts multi-beam sonar images into a range representation spatially corresponding to RGB image, which extends acoustic information from sparse range constraints to cross-modal feature fusion. This dual-branch network extracts and fuses visual texture features and sonar scale features to predict a dense metric depth. These visual-acoustic fusion methods have evolved from global scale correction, local depth correction, and training constraints to cross-modal alignment and feature fusion. Acoustic information provides absolute-range references for monocular vision, while visual information recovers continuous scene structures. Their combination alleviates monocular scale ambiguity, although the results remain affected by sonar resolution, sensor-calibration errors, and cross-modal spatial correspondence.

Visual-Laser Fusion: Visual-laser fusion observes projected laser spots or lines on underwater surfaces, and uses triangulation to calculate scene depth. Active laser patterns do not depend on object texture and can provide denser depth points with actual scale than sparse acoustic ranges. Existing related methods can be divided into single-frame local depth measurement and continuous depth measurement. Campos and Codina [49] combined a monocular camera with a laser projector, where the camera detects projected patterns and its geometry related to the projector yields local depth. This active projection provides identifiable measurement locations in textureless regions. Inglis *et al.* [50] observed a projected laser line, triangulated a depth profile at each instant, and transformed profiles into a common coordinate system using vehicle motion. Bodenmann *et al.* [51] used one camera and a line laser to acquire depth map and image color to reconstructed points, then sea trials demonstrated continuous metric-scale recovery in deep water environments. Visual-laser fusion methods provide metric depth, and are effective for precise near-field measurement in weakly textured areas. But their coverages are limited by the projected laser and

underwater absorption or scattering, making dense depth estimation difficult.

Visual-SLAM Fusion: Unlike acoustic and laser sensors, SLAM does not measure range directly, and recovers camera motion and sparse map points from image sequences, and these cross-frame geometric outputs constrain depth prediction. Visual-SLAM fusion refers to information-level fusion using SLAM poses, map points or scale. Ebner *et al.* [4] proposed TRUDepth, which feeds RGB images together with sparse depth priors into a network, uses sparse depth to provide scale constraints, and combines image features to predict dense depth of the entire image, thereby solving the problem that sparse map points are difficult to form a complete depth map. On the FLSea dataset, this method matches image features of adjacent frames and extracts depths at corresponding positions from ground-truth depth map to simulate sparse priors, and on the Lizard Island dataset, map points generated by underwater SLAM are used as input. Experimental results show that incorporating sparse depth priors can reduce underwater depth prediction error and improve the accuracy of predicted scale. Besides, Zhang *et al.* [52] proposed the Sparsity Adaptive Depth Estimator (SPADE), which first uses a pretrained monocular model to obtain relative depth, then performs global scale alignment based on sparse depth points generated by visual-inertial SLAM, and subsequently generates dense metric depth via a scale-refinement network. Unlike TRUDepth, which directly feeds RGB images and sparse depth into the network together, SPADE first corrects the global scale of relative depth and then refines the scale deviations at different locations, and is able to adapt to variations in the number and distribution of sparse depth points. These visual-SLAM fusion methods mainly use sparse geometric information from image sequences to constrain dense depth prediction. Existing methods have evolved from direct fusion of RGB images and sparse depth to relative-depth scale alignment and pixel-wise refinement. Such methods mitigate monocular scale ambiguity, but inaccurate feature matching or pose estimation

under underwater degradation can produce erroneous sparse depth and degrade final results.

Overall, multi-modal fusion is intended to address scale ambiguity and depth instability in monocular underwater vision. Among these, visual–acoustic fusion targets global scale and long-range perception, visual–laser fusion supplements local depth observations and improve the recovery of close-range details and geometric structures, and visual–SLAM fusion emphasizes temporal consistency and global scale during continuous motion. Together, the three constitute an important route in underwater depth estimation, advancing from single-frame visual inference toward collaborative multi-source perception.

4 Representative Applications of Underwater Depth Estimation

Underwater depth estimation recovers scene range distribution, and provides spatial information for diverse applications. This section presents representative applications of underwater depth estimation, including underwater image enhancement and restoration, underwater object detection, AUV navigation and localization, underwater 3D reconstruction, and underwater robotic operations.

4.1 Underwater Image Enhancement and Restoration

Underwater image enhancement and restoration is one of the most direct applications of underwater depth estimation. Because of light absorption and scattering in water, underwater images often suffer from color casts, low contrast, and blurring, and the severity of these degradations varies with object range. Depth can describe light-propagation distance in different regions and assist transmission estimation, backscatter correction, and color recovery, thereby improving underwater image enhancement and restoration^[53, 54]. The development of underwater depth estimation for underwater image enhancement and restoration is illustrated in Fig. 4.

Early work focused on inferring depth from color and intensity. Bianco *et al.*^[55] estimated scene depth from inter-channel intensity differences and used it to restore single underwater image. Chiang and Chen^[56] compensated color channels according to wavelength-dependent attenuation. Drews *et al.*^[57] introduced the UDCP, using green and blue channels to estimate transmission and connect depth inference with image restoration. Further, Drews *et al.*^[13] validated statistical assumptions to generate relative-depth maps, and restored images from estimated medium transmissions. These approaches obtained depth estimation from attenuation or transmission indirectly. Beyond color attenuation cues, depth information have been extracted from other forms of underwater image degradations. Peng and Cosman^[11] proposed the image blurriness and light absorption (IBLA) method, which combines image blurriness with light absorption to estimate scene depth and restores different regions according to range variation. Akkaynak and Treibitz^[30] re-examined underwater image formation, and showed that direct-signal attenuation and backscatter should be described using different parameters. This revision clarifies different effects of scene depth on two degradation components. Besides, Akkaynak and Treibitz^[58] proposed Sea-thru, which obtains scene

depth from multi-view images, and estimates direct-signal attenuation and backscatter parameters to recover underwater object colors. WaterGAN synthesized underwater images from in-air RGB-D data, using depth maps to control color attenuation and scattering in different regions, thereby providing training data for unsupervised color correction^[59]. It demonstrates that depth can serve as a condition for generating underwater degradation. Ye *et al.*^[8] placed underwater depth estimation and color correction in a unified framework, and recovered scene depth and image color simultaneously to share degradation information via unsupervised adaptation. PUGAN^[60] combined an underwater image-formation model with an adversarial network, and used physical parameters and degradation information to guide underwater image enhancement. UWdepth^[33] performed unsupervised depth estimation from consecutive underwater images, and inserted the predicted depth into the revised underwater image-formation model^[30] to enhance each region according to its propagation distance. WaterMono^[47] established a bidirectional relationship between depth estimation and underwater image enhancement, where predicted depth guides image enhancement, while enhanced images improve brightness and structural consistency to constrain stable training for unsupervised depth estimation. Xiao *et al.*^[61] introduced depth-map consistency into a semi-supervised dual-cycle network to perform underwater image enhancement and depth estimation simultaneously. Du *et al.*^[41] further incorporated underwater image enhancement and depth estimation into a unified physical-model-guided framework, which jointly estimates scene depth, backscatter and optical parameters, and constrains training by re-degrading restored images.

For underwater image enhancement and restoration, depth is used in three main ways: inferring scene range from color, transmission or blur, computing degradation parameters from measured or predicted depth, and jointly training depth and color recovery. These strategies adjust restoration strength to depth estimation, although range errors propagate into color compensation and backscatter correction.

4.2 Underwater 3D Scene Reconstruction

Underwater depth estimation provides the distance between different image locations and the camera, allowing scene structures in 2D images to be recovered in 3D space. Depth obtained from single image corresponds only to current viewpoint, complete scene reconstruction must combine camera poses and multi-view correspondences to align depth from different viewpoints in a common coordinate system. Representative works of underwater depth estimation for underwater three-dimensional scene reconstruction is summarized in Fig. 5.

Wu *et al.*^[1] proposed SDU-SfM, incorporating underwater depth into 3D reconstruction. Relative depth is first estimated from each view through nonlinear domain projection, and then supplement scene points to guide image restoration and feature filtering. The accuracy of dense depth directly affects scene structure recovery. Yu *et al.*^[15] introduced UDepth to provide relatively dense range for underwater 3D reconstruction, which uses underwater light-attenuation laws to construct a domain-projection constraint and predicts pixel-wise depth from a single underwater image. To address the lack of metric scale in monocular depth, Ebner *et al.*^[4]

combined RGB images with sparse depth from triangulated feature points, using sparse depth to constrain global scale and image to complete remaining regions. Underwater image quality also affects depth prediction and cross-view matching. Varghese *et al.* [20] presented USe-ReDI-Net, which combines underwater image-formation model with inter-frame geometry to recover both scene depth and clear image from single frame. This method outputs remain depth maps and restored images, providing clearer imagery and corresponding range information for underwater 3D reconstruction. Wang *et al.* [62] designed UW-GS, which first obtains pseudo

depth for input images using an existing monocular depth model, and then uses this range to determine spatial positions of 3D Gaussians, reducing reconstruction bias caused by underwater degradations. This work shows that monocular depth can provide auxiliary range information for underwater 3D reconstruction. Additionally, Yang *et al.* [63] developed UWGS-SLAM to construct 3D Gaussian scene representation from RGB images, then estimated depth maps and integrated underwater medium modeling with camera tracking and scene reconstruction. It demonstrates that depth estimation can initialize spatial representation and contribute to localization

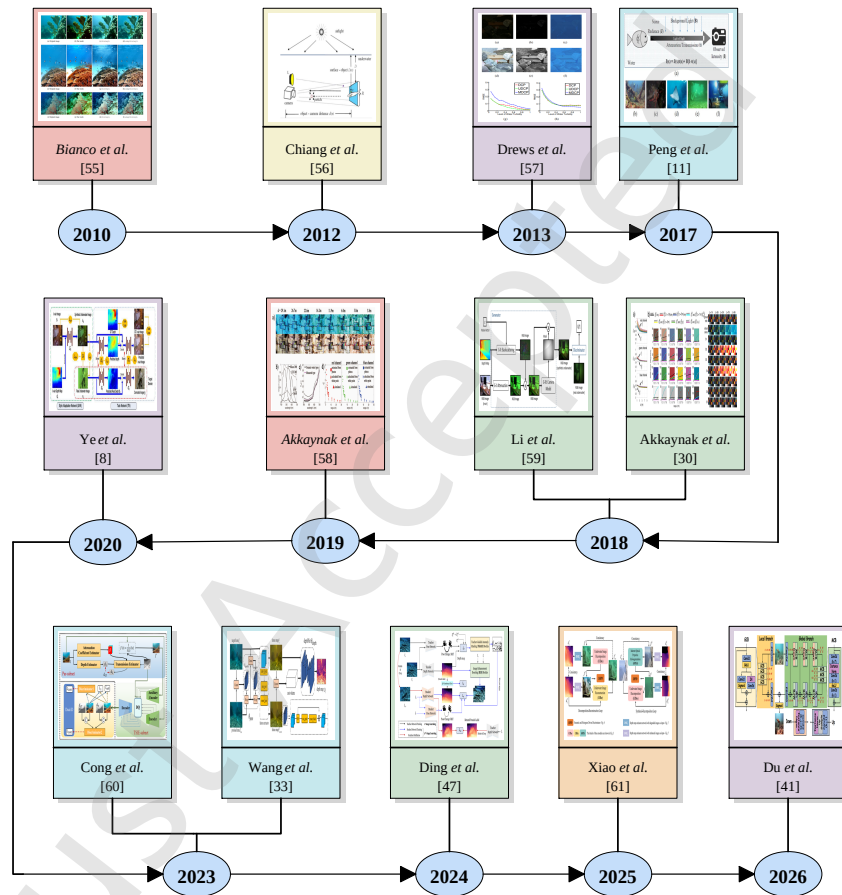


Fig. 4 Representative works of underwater image enhancement and restoration with depth estimation.

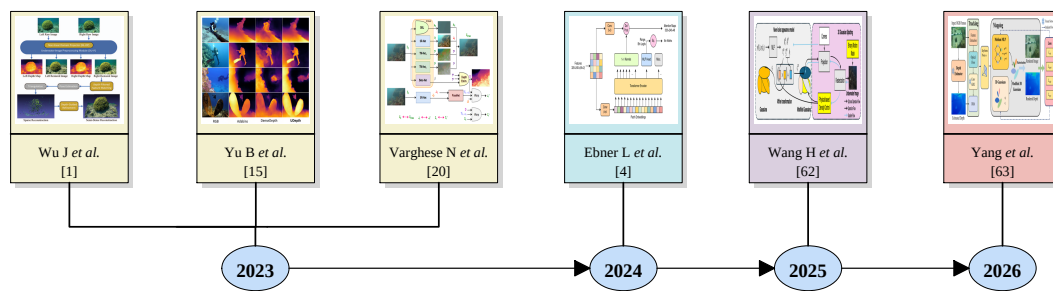


Fig. 5 Representative works of underwater 3D reconstruction with depth estimation.

and dense reconstruction under water.

Underwater depth estimation contributes pixel-wise range, complements sparse structure, and constrains 3D representations. Since monocular estimates can have scale ambiguity and local errors, stable reconstruction still requires camera poses, multi-view correspondence, or sparse metric depth.

4.3 Underwater Object Detection

Underwater object detection requires targets to be identified and localized in images. Since targets appear at different scales and levels of clarity at different ranges, relying on color, texture and contour is susceptible to underwater image [64]. Depth can reflect the distance between different image regions and the camera, providing a basis for object discrimination. Representative works of underwater depth information for underwater object detection are shown in Fig. 6.

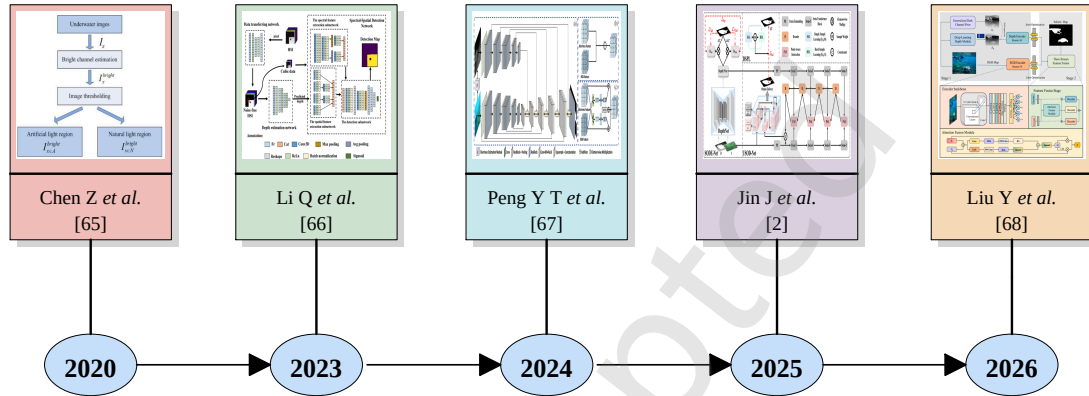


Fig. 6 Representative works of underwater object detection with depth estimation.

Chen *et al.* [65] extracted 3D depth features from a single underwater image and combined them with image color and brightness, and depth features add spatial information to underwater object detection. To address noise in underwater depth maps, Jin *et al.* [2] proposed target-depth enhancement, which increases depth difference between target and background while maintaining smoothness within each region, thereby improving target localization accuracy. Li *et al.* [66] predicted target range with a depth-estimation network, and selected corresponding detector according to that prediction, enabling detection to accommodate distance-dependent changes in image characteristics. When reliable depth maps are unavailable, distance-dependent appearance changes serve as auxiliary cues. Peng *et al.* [67] exploited progressive blurring of underwater targets with distance, extracted input blurriness to combine with RGB images for object detection. Liu *et al.* [68] further integrated underwater monocular depth estimation and underwater object detection within one network. Its joint training encourages predicted depth to retain information about object positions and boundaries, while detection results provide feedback to depth estimation.

Depth can distinguish target from background, and support object localization and boundary delineation. These methods fuse depth features with spatial information, exploit distance-dependent image changes, or jointly train underwater depth estimation and object detection. Therefore, obtaining stable depth information and preventing depth errors from propagating into detection remain essential for depth-assisted underwater object detection.

4.4 AUV Navigation and Localization

During underwater navigation, an AUV determines its position and the distances from the vehicle to obstacles or seafloor.

Depth estimation obtains scene-range from images, and supports obstacle avoidance and path planning. Acoustic ranging, inertial data, and camera poses further constrain depth scale, and maintain consistent and continuous measurements. Representative works of underwater depth estimation for AUV Navigation and Localization is illustrated in Fig. 7.

Early studies used relative-distance cues from underwater monocular images to assist navigation, and later progressively incorporated inertial, pressure, and acoustic-ranging to improve localization and scale estimation. Gaya *et al.* [69] used the CNN to estimate transmission maps from monocular underwater images, derived relative depth from the relationship between transmission and distance, and selected open regions as obstacle-avoidance directions. Ferrera *et al.* [70] released the AQUALOC dataset, which provides underwater images, inertial measurements, and pressure data as the test basis for visual localization and SLAM near the seafloor. Rozner *et al.* [5] combined a single-beam echo sounder, a monocular camera and ORB-SLAM2, where acoustic ranging supplied metric distance, while visual feature tracking propagated scale constraint to camera trajectory and map points, thereby reducing monocular scale drift. As visual perception increasingly involved in motion planning, AUVs began to use scene information to adjust navigation strategies. AquaVis [3] incorporated target visibility into planning, and thus the AUV could reach its goal and avoid obstacles while retaining observations of important objects. Yang *et al.* [71] introduced estimated depth, single-beam sonar, target position, and action history into a navigation policy for collision avoidance in unknown environments, where depth describes forward structure and sonar supplies metric range. Later work emphasized deployable underwater monocular depth estimation. Amitai *et al.* [19] used consecutive-frame reprojection for self-supervised training. Ebner *et al.* [4] fused

sparse triangulated depth with RGB images to generate dense metric depth and reduce scale error. Besides, UDepth^[15] predicts depth from one image by a lightweight network, enabling low-cost robots to perceive range in real-time. Recent work has combined depth and terrain information with multi-sensor localization systems. NeuRSS^[72] used side-scan sonar to estimate seabed elevation as a terrain constraint for SLAM, and then improved poses were used to update bathymetry results, allowing underwater localization and depth estimation to correct each other. AQUA-SLAM^[73] fused stereo cameras and a Doppler velocity log to improve pose-estimation accuracy during continuous underwater navigation. AquaticVision^[74] provided event-camera, frame-camera, and ground-truth trajectory data for evaluating visual localization under underwater low-light and blurred conditions. Further, SA-MPC^[75] fused monocular depth estimation with single-beam echo-sounder to produce high-resolution depth maps, and incorporated environmental safety into model control for dynamic-obstacle perception and avoidance in complex water.

Depth estimation can provide obstacle range and forward structure, while acoustic, inertial and SLAM information can correct scale and maintain temporal geometry. More accurate and reliable depth can improve avoidance, path planning, and near-bottom navigation under water.

4.5 Underwater Robotic Operations

When performing docking, grasping, and equipment operations, underwater robots determine spatial position of targets and their distances. Underwater monocular depth estimation can recover range information for targets and surrounding regions from images, supporting target approach, pose adjustment, and manipulator-motion planning. Representative research of underwater depth estimation for underwater robotic operations is shown in Fig. 8.

Yu *et al.*^[15] proposed UDepth to construct depth cues from underwater light attenuation, and utilized a lightweight network combining MobileNetV2 and a Transformer to predict dense depth from a single underwater image. This method can run on embedded devices, allowing underwater robots to obtain scene range continuously for real-time target approach and motion adjustment. Ebner *et al.*^[4] fed underwater monocular image and sparse depth priors into a network, and used range information from triangulated features to correct monocular scale ambiguity. It improves depth accuracy at close range, and supports range assessment for underwater intervention and close-range inspection. To reduce the effect of underwater image degradation on depth prediction, Jiang *et al.*^[76] used dark channel prior to compute the transmission map containing depth information, and constrained background predictions with light attenuation prior. This lightweight network balances depth accuracy and computational efficiency, facilitating deployment on resource-constrained underwater platforms. Further, CD-UDepth^[36] extracted depth features from RGB image and light-attenuation separately, and fused them according to their confidence, improving depth predictions in regions with color distortion, blurred textures, and long

range. Sonar-MASt3R^[77] fused dense visual correspondences from an end-effector camera with the geometry and metric scale supplied by imaging sonar, and thus reconstructed a metric 3D workspace in real-time under turbid and unstructured water.

Underwater monocular depth estimation for robotic operations provide accurate metric scale, fast execution, and stable results in complex underwater conditions. These above-mentioned depth estimation methods provide robots with scene range and spatial structure, supporting target approach, operation planning, and close-range manipulation.

5 Challenges and Future Directions

5.1 Major Challenges

Underwater depth is commonly estimated from color, texture, and contours of objects in an underwater image, however, these features are easily affected by light absorption, scattering, and illumination changes in water, which cannot reflect reliable scene depth. Additionally, the scarcity of real depth labels limits existing models training, while high computational cost makes it difficult to balance depth estimation accuracy and processing efficiency. The principal challenges of underwater depth estimation are summarized in Table 5.

Complex underwater image degradation: Underwater images are usually affected by light absorption and scattering. Wavelength-dependent attenuation causes color casts, reduced contrast, and blurred details, while background light or suspended particles alter image brightness. These degradations are related to propagation distance but vary with water types and non-uniform illuminations, preventing stable correspondence between image appearance and scene depth. Attenuation of target-reflected light and backscatter should be represented by different parameters^[30], and applying atmospheric dehazing models cannot describe underwater image formation accurately. Moreover, scene range is used to estimate two degradation components separately, indicating that depth can explain underwater image changes when considered together with water optical properties^[58]. Existing methods relying on color, brightness, or texture are therefore prone to depth errors when water quality or illumination changes.

Insufficient texture and illumination interference: Seafloor sand, rock surfaces, and turbid backgrounds typically lack distinctive textures, while scattering weakens underwater object boundaries and deprives monocular images of stable spatial cues. Local highlights, shadows, and nonuniform brightness caused by artificial lighting may be misinterpreted as the changes in scene range. Moving organisms and suspended particles also disrupt pixel correspondences between frame, and weaken reconstruction constraints in existing unsupervised methods. It is noted that image changes caused by underwater medium reduce monocular depth accuracy and therefore require unsupervised training process to be adapted to underwater image formation^[19]. Besides, extracting reliable depth cues from regions with weak texture and unstable illumination^[47] remains an open problem.

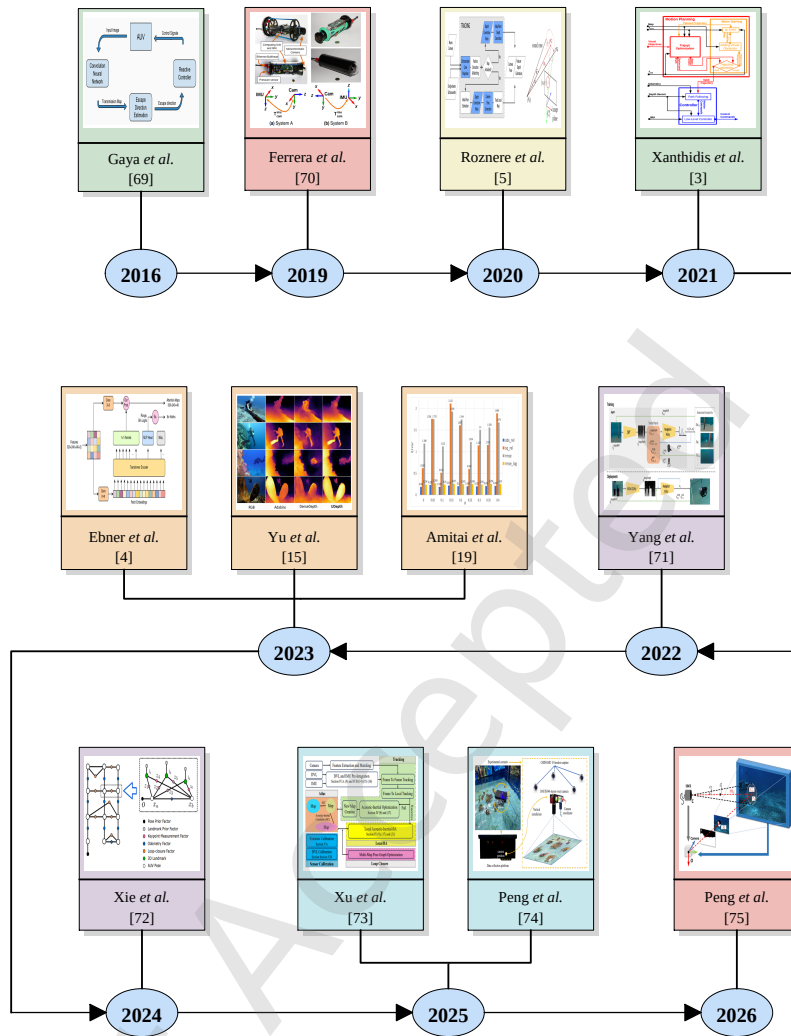


Fig. 7 Representative works of AUV navigation and localization with depth estimation.

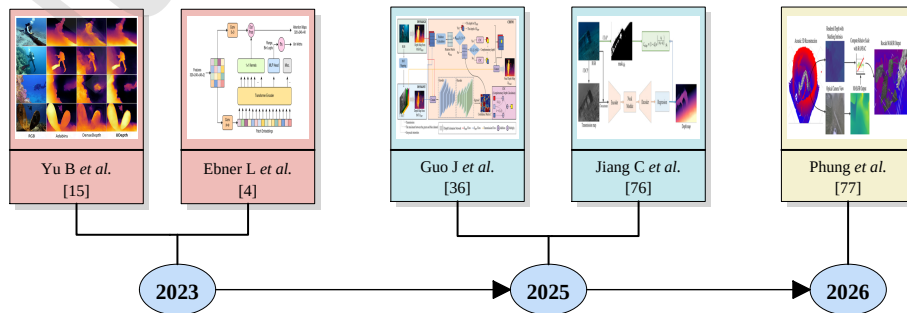


Fig. 8 Representative works of underwater robotic operations with depth estimation.

Limited cross-domain generalization: Water bodies differ substantially in turbidity, color attenuation, suspended particles, and non-uniform illumination. When a model trained on a particular dataset or test pool is applied to another water body, camera or scene, input distribution changes and depth-estimation performance decline. Underwater imaging characteristics are usually incorporated into monocular depth estimation to reduce errors by transferring terrestrial models to underwater scenes^[34]. The evaluation^[78] shows that terrestrial vision foundation models still exhibit substantial domain shift in underwater images because of light attenuation, scattering, and color casts. Adaptation to different water bodies and imaging devices is therefore essential for moving from dataset testing to practical deployment.

Tradeoff of accuracy and efficiency: Transformers and hybrid networks use broad image context, but require relatively high computation. AUV, remotely operated vehicles (ROVs) or other underwater robots are constrained by processor performance, power and storage, and depth estimation models therefore should retain accuracy while meeting speed requirements of continuous underwater navigation and real-time operations. The lightweight backbones, such as Mamba^[35] or Tree-Mamba^[18], with approximately linear computational complexity, can be designed with attenuation priors^[15] or sparse priors^[4], to maintain underwater depth estimation performance while reducing computational cost.

Scarcity of high-quality real data: Supervised learning methods require a large amount of underwater images paired with ground-truth depth, but underwater depth ground truths are difficult to be acquired. Laser, sonar, and stereo systems are affected by water scattering, measurement range, and sensor calibration, while equipment cost and acquisition difficulty limit dataset scale. Existing datasets concentrate on a few types of water bodies and scenes, and not represent the diversity of real water environments fully. The available datasets, such as FLSea^[79] or BlueDepth^[18], remain insufficient in data volume, scene coverage, and ground-truth accuracy, potentially biasing model training and cross-dataset evaluation.

5.2 Future Directions

In response to the challenges discussed above, future directions for underwater depth estimation can be outlined as follows:

Integration of underwater imaging priors and deep learning: Light absorption and scattering in water alter image color, brightness and sharpness, while the relationship between these changes and scene depth depends on water conditions. The physically guided Transformer, or multi-physics self-supervised methods have incorporated underwater attenuation, medium transmission, and image-formation models into feature extraction or training^[34, 41, 39]. Future studies should reduce reliance on fixed optical parameters, enabling deep learning-based models to infer current attenuation and scattering conditions from inputs and adjust depth estimations accordingly. Consistency between underwater image restoration and re-degradation can also be used to validate predictions, and improve depth estimation at turbid water, low-light, and artificial illumination conditions.

Improvement of multimodal fusion reliability: Information from different sources can complement one another under complex underwater conditions while limiting the propagation

of their respective errors into depth predictions. Existing works use sparse depth information to correct the scale of monocular predictions^[52] and camera-sonar extrinsic calibration to determine the relative position and orientation of two sensors^[80]. But using such information for depth estimation requires identifying image locations corresponding to the measurements and assessing the reliability of different sources when their observations are inconsistent. Future research should improve the accuracy of correspondences between underwater sensor observations, and determine the contribution of each source to depth estimation based on factors such as image clarity and measurement error, thereby reducing the influence of inaccurate range or positional constraints on depth prediction.

Construction of real-world underwater depth datasets and model evaluation: Real underwater images and their corresponding scene-depth maps are fundamental to model training and evaluation. Light absorption and scattering, however, make large-area dense-depth acquisition difficult, and existing datasets also differ in label source, accuracy, and applicable scenarios. Future dataset construction should focus on improving the reliability of depth labels, covering different water qualities, observation distances, illumination conditions, and dynamic conditions, while clearly specifying calibration methods and error ranges. In addition to commonly used depth errors, model evaluation should examine real-world scale, low-visibility regions, temporal stability across consecutive frames, and cross-water-domain performance.

Real-time depth estimation for underwater intelligent devices: Underwater depth models need to run continuously on AUV, ROVs or underwater robots, whose processor performance, storage space, and power consumption are typically constrained. For example, UDepth^[15] adopted a lightweight network to accelerate depth prediction speed, and Tree-Mamba^[18] reduced the cost of long-range feature modeling while improving the representation of underwater structures. Future research should evaluate not only parameter counts and single-frame speed, but should also examine end-to-end latency, memory usage, power consumption, and long-term stability on actual underwater intelligent devices. For underwater navigation, obstacle avoidance, and precise operations, emphasis should be placed on evaluating metric depth errors near obstacles and at target boundaries, and on enabling models with mechanisms to identify unreliable predictions under low-light, turbid, or occluded conditions.

6 Conclusion

This article systematically surveys the study and progress of underwater depth estimation. At the theoretical level, underwater image-formation models, light-attenuation laws, Bayesian estimation, backscattering, and visual priors clarify the intrinsic relationship between underwater image degradation and scene depth. At the methodological level, traditional physical priors, deep-learning-driven, and multimodal fusion approaches are summarized, and their cores, advantages and limitations are analyzed. At the application level, the role of depth estimation in image enhancement, 3D reconstruction, object detection, AUV navigation and localization, and robotic operations under water is presented. Current research remains challenged faced by complex image degradation, weak texture and

Table 5 Major challenges in underwater depth estimation

Challenge	Manifestation	Literature
Complex underwater image degradation	Color cast, scattering, backscatter, and nonuniform illumination affect imagery	Akkaynak and Treibitz ^[30] Akkaynak and Treibitz ^[58]
Insufficient texture and illumination interference	Weak textures, blurred edges, local highlights, and dynamic objects corrupt depth cues	Amitai <i>et al.</i> ^[19] Ding <i>et al.</i> ^[47]
Limited cross-domain generalization	Low performance across waters, datasets and sensors	Wu <i>et al.</i> ^[34] Cai and Metzler ^[78]
Tradeoff of accuracy and efficiency	Accurate models are difficult to run on AUV/ROV in real-time	Yu <i>et al.</i> ^[15] Ebner <i>et al.</i> ^[4]
Scarcity of high-quality real data	Ground truth is difficult to obtain and datasets cover limited scenes	Randall <i>et al.</i> ^[79] Zhuang <i>et al.</i> ^[18]

illumination interference, limited cross-water generalization, tradeoff between accuracy and efficiency, and scarce real-world depth data. Additionally, future research should integrate underwater imaging priors with deep learning-based methods, improve the reliability of multimodal fusion, enhance real-world underwater depth datasets and model evaluation, and develop real-time depth estimation for autonomous underwater devices, thereby improving the robustness and practical applicability.

Underwater depth estimation has been a foundation of underwater image sensing and understanding, and is a key link among image-formation physics, 3D perception, and intelligent operation under water. As the demand grows for ocean exploration, AUV navigation, and underwater robotic intervention, the technique of underwater depth estimation will advance underwater visual perception technology toward more intelligent and reliable solutions.

7 Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 62171252, 61701245; in part by the Fundamental Research Funds for the Central Universities under Grant 00007764.

References

- [1] J. Wu, B. Yu, and M. J. Islam, "3d reconstruction of underwater scenes using nonlinear domain projection," in *IEEE Conference on Artificial Intelligence*, 2023, pp. 359–361.
- [2] J. Jin, Q. Jiang, Q. Wu, B. Xu, and R. Cong, "Underwater salient object detection via dual-stage self-paced learning and depth emphasis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 3, pp. 2147–2160, 2025.
- [3] M. Xanthidis, M. Kalaitzakis, N. Karapetyan, J. Johnson, N. Vitzilaos, J. M. O’Kane, and I. Rekleitis, "Aquavis: A perception-aware autonomous navigation framework for underwater vehicles," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021, pp. 5410–5417.
- [4] L. Ebner, G. Billings, and S. Williams, "Metrically scaled monocular depth estimation through sparse priors for underwater robots," in *IEEE International Conference on Robotics and Automation*, 2024, pp. 3751–3757.
- [5] M. Roznere and A. Q. Li, "Underwater monocular image depth estimation using single-beam echosounder," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020, pp. 1785–1790.
- [6] T. Dolereit, U. F. von Lukas, and A. Kuijper, "Underwater stereo calibration utilizing virtual object points," in *IEEE OCEANS*, 2015, pp. 1–7.
- [7] M. Roznere, A. K. Pediredla, S. E. Lensgraf, Y. Girdhar, and A. Q. Li, "Underwater dome-port camera calibration: Modeling of refraction and offset through n-sphere camera model," in *IEEE International Conference on Robotics and Automation*, 2024, pp. 6110–6117.
- [8] X. Ye, Z. Li, B. Sun, Z. Wang, R. Xu, H. Li, and X. Fan, "Deep joint depth estimation and color correction from monocular underwater images based on unsupervised adaptation networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3995–4008, 2020.
- [9] Q. Cheng, Z. Wang, X. Mu, and W. Wu, "Cnn-transformer feature aggregation for underwater self-supervised multiframe monocular depth estimation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–12, 2025.
- [10] J. Zhou, J. Sun, C. Li, Q. Jiang, M. Zhou, K. Lam, W. Zhang, and X. Fu, "Hclr-net: Hybrid contrastive learning regularization with locally randomized perturbation for underwater image enhancement," *International Journal of Computer Vision*, vol. 132, no. 10, pp. 4132–4156, 2024.
- [11] Y.-T. Peng and P. C. Cosman, "Underwater image restoration based on image blurriness and light absorption," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1579–1594, 2017.

- [12] Y.-T. Peng, X. Zhao, and P. C. Cosman, "Single underwater image enhancement using depth estimation based on blurriness," in *IEEE International Conference on Image Processing*, 2015, pp. 4952–4956.
- [13] P. L. Drews, E. R. Nascimento, S. S. Botelho, and M. F. Montenegro Campos, "Underwater depth estimation and image restoration based on single images," *IEEE Computer Graphics and Applications*, vol. 36, no. 2, pp. 24–35, 2016.
- [14] E. Silva Vaz, E. F. de Toledo, and P. L. Drews, "Underwater depth estimation based on water classification using monocular image," in *Latin American Robotics Symposium (LARS), 2020 Brazilian Symposium on Robotics and 2020 Workshop on Robotics in Education*, 2020, pp. 1–6.
- [15] B. Yu, J. Wu, and M. J. Islam, "Udepth: Fast monocular depth estimation for visually-guided underwater robots," in *IEEE International Conference on Robotics and Automation*, 2023, pp. 3116–3123.
- [16] G. Yang, G. Kang, J. Lee, and Y. Cho, "Joint-id: Transformer-based joint image enhancement and depth estimation for underwater environments," *IEEE Sensors Journal*, vol. 24, no. 3, pp. 3113–3122, 2024.
- [17] C. Wang, H. Xu, G. Jiang, M. Yu, T. Luo, and Y. Chen, "Underwater monocular depth estimation based on physical-guided transformer," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.
- [18] P. Zhuang, Y. Wang, Z. Fu, H. Zhang, S. Kwong, and C. Li, "Tree-mamba: A tree-aware mamba for underwater monocular depth estimation," *arXiv preprint arXiv:2507.07687*, 2025.
- [19] S. Amitai, I. Klein, and T. Treibitz, "Self-supervised monocular depth underwater," in *IEEE International Conference on Robotics and Automation*, 2023, pp. 1098–1104.
- [20] N. Varghese, A. Kumar, and A. N. Rajagopalan, "Self-supervised monocular underwater depth recovery, image restoration, and a real-sea video dataset," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 214–12 224.
- [21] J. Zhou, D. Zhang, Z. He, Q. Gai, and Q. Jiang, "Multi-prior fusion transfer plugin for adapting in-air models to underwater image enhancement and detection," *IEEE Transactions on Image Processing*, vol. 34, pp. 7773–7785, 2025.
- [22] J. Zhou, C. Liu, Q. Jiang, X. Fu, J. Hou, and X. Li, "Semantic contrast for domain-robust underwater image quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 6, pp. 6279–6296, 2026.
- [23] K. Mantani, T. Katayama, T. Song, and T. Shimamoto, "Depth estimation with sonar-based correction for low-cost underwater drone," in *IEEE OCEANS*, 2022, pp. 1–4.
- [24] H. Liu, M. Roznere, and A. Q. Li, "Deep underwater monocular depth estimation with single-beam echosounder," in *IEEE International Conference on Robotics and Automation*, 2023, pp. 1090–1097.
- [25] P. Lv, Y. Liu, Z. Jiang, S. Chai, F. Zha, P. Wang, and W. Guo, "An rgb-sonar fusion framework for underwater depth estimation enabled by cross-modal alignment," *Optics and Lasers in Engineering*, vol. 205, p. 109879, 2026.
- [26] J. Jiang, Z. Li, and X. Liu, "Deep learning based monocular depth estimation: A survey," *Chinese Journal of Computers*, vol. 45, no. 6, pp. 1276–1307, 2022.
- [27] P. Zhuang, J. Wu, F. Porikli, and C. Li, "Underwater image enhancement with hyper-laplacian reflectance priors," *IEEE Transactions on Image Processing*, vol. 31, pp. 5442–5455, 2022.
- [28] X. Yang, X. Zhang, N. Wang, G. Xin, and W. Hu, "Underwater self-supervised depth estimation," *Neuro-computing*, vol. 514, pp. 362–373, 2022.
- [29] D. F. Swinehart, "The beer-lambert law," *Journal of chemical education*, vol. 39, no. 7, p. 333, 1962.
- [30] D. Akkaynak and T. Treibitz, "A revised underwater image formation model," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6723–6732.
- [31] A. Halimi, A. Maccarone, A. McCarthy, S. McLaughlin, and G. S. Buller, "Object depth profile and reflectivity restoration from sparse single-photon data acquired in underwater environments," *IEEE Transactions on Computational Imaging*, vol. 3, no. 3, pp. 472–484, 2017.
- [32] A. Halimi, A. Maccarone, R. A. Lamb, G. S. Buller, and S. McLaughlin, "Robust and guided bayesian reconstruction of single-photon 3d lidar data: Application to multi-spectral and underwater imaging," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 961–974, 2021.
- [33] J. Wang, X. Ye, Y. Liu, X. Mei, and J. Hou, "Underwater self-supervised monocular depth estimation and its application in image enhancement," *Engineering Applications of Artificial Intelligence*, vol. 120, p. 105846, 2023.
- [34] X. Wu, J. Wang, J. Wang, S. Rong, and B. He, "Umono: Physical-model-informed hybrid cnn–transformer framework for underwater monocular depth estimation," *IEEE Journal of Oceanic Engineering*, vol. 51, no. 1, pp. 780–793, 2026.
- [35] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
- [36] J. Guo, J. Ma, F. Sun, Z. Gao, Ángel F. García-Fernández, H.-N. Liang, X. Zhu, and W. Ding, "Cd-udepth: Complementary dual-source information fusion for underwater monocular depth estimation," *Information Fusion*, vol. 118, p. 102961, 2025.

- [37] J. Zhou, Q. Liu, Q. Jiang, W. Ren, K.-M. Lam, and W. Zhang, "Underwater camera: Improving visual perception via adaptive dark pixel prior and color correction," *International Journal of Computer Vision*, vol. 133, no. 11, pp. 8215–8233, 2025.
- [38] H. Tamaki, T. Katayama, T. Song, and T. Shimamoto, "Color correction method using monocular depth estimation model for underwater images," in *IEEE 13th Global Conference on Consumer Electronics*, 2024, pp. 641–643.
- [39] F. Xiao and J. Qu, "Self-supervised underwater monocular depth estimation informed by multi-physics processes," *IEEE Robotics and Automation Letters*, vol. 10, no. 9, pp. 9144–9151, 2025.
- [40] S. González-Sabbagh, A. Robles-Kelly, and S. Gao, "Scene-cgan: A gan for underwater restoration and scene depth estimation," *Computer Vision and Image Understanding*, vol. 250, p. 104225, 2025.
- [41] D. Du, L. Si, F. Xu, J. Niu, and F. Sun, "A physical model-guided framework for underwater image enhancement and depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 36, no. 7, pp. 10 244–10 259, 2026.
- [42] W. Chen, X. Luo, F. Li, and D. Wang, "Estimation of underwater monocular depth using laplacian pyramid-based depth residuals," in *International Conference on Information Science and Technology*, 2023, pp. 40–47.
- [43] G. Yang and Y. Cho, "Trident: Efficient triple-task learning of dehazing, depth, and uncertainty estimation for underwater 3-d robot visual perception," *IEEE Sensors Journal*, vol. 25, no. 1, pp. 858–870, 2025.
- [44] H. Zhu and Z. Xing, "Udepthdiff: Multi-modal physical diffusion-based underwater monocular depth estimation," *IEEE Access*, vol. 14, pp. 26 953–26 964, 2026.
- [45] H. Gupta and K. Mitra, "Unsupervised single image underwater depth estimation," in *IEEE International Conference on Image Processing*, 2019, pp. 624–628.
- [46] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3827–3837.
- [47] Y. Ding, K. Li, H. Mei, S. Liu, and G. Hou, "Watermono: Teacher-guided anomaly masking and enhancement boosting for robust underwater self-supervised monocular depth estimation," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–14, 2025.
- [48] D. Liu, Y. Liu, B. Li, and Q. Duan, "Enhanced depth estimation for underwater monocular images via intrinsic scale invariance and skip expanded context fusion," *Applied Soft Computing*, vol. 201, p. 115560, 2026.
- [49] M. Massot-Campos and G. Oliver-Codina, "Underwater laser-based structured light system for one-shot 3d reconstruction," in *IEEE SENSORS*, 2014, pp. 1138–1141.
- [50] G. Inglis, C. Smart, I. Vaughn, and C. Roman, "A pipeline for structured light bathymetric mapping," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 4425–4432.
- [51] A. Bodenmann, B. Thornton, and T. Ura, "Generation of high-resolution three-dimensional reconstructions of the seafloor in color using a single camera and structured light," *Journal of Field Robotics*, vol. 34, no. 5, pp. 833–851, 2017.
- [52] H. Zhang, G. Billings, and S. B. Williams, "Spade: Sparsity adaptive depth estimator for zero-shot, real-time, monocular depth estimation in underwater environments," *IEEE Journal of Oceanic Engineering*, pp. 1–11, 2026.
- [53] H. Wang, S. Sun, and P. Ren, "Underwater color disparities: Cues for enhancing underwater images toward natural color consistencies," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 2, pp. 738–753, 2024.
- [54] H. Wang, W. Zhang, L. Bai, and P. Ren, "Metalantis: A comprehensive underwater image enhancement framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–19, 2024.
- [55] N. Carlevaris-Bianco, A. Mohan, and R. M. Eustice, "Initial results in underwater single image dehazing," in *IEEE OCEANS*, 2010, pp. 1–8.
- [56] J. Y. Chiang and Y.-C. Chen, "Underwater image enhancement by wavelength compensation and dehazing," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1756–1769, 2012.
- [57] P. Drews Jr, E. do Nascimento, F. Moraes, S. Botelho, and M. Campos, "Transmission estimation in underwater single images," in *IEEE International Conference on Computer Vision Workshops*, 2013, pp. 825–830.
- [58] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1682–1691.
- [59] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 387–394, 2018.
- [60] R. Cong, W. Yang, W. Zhang, C. Li, C.-L. Guo, Q. Huang, and S. Kwong, "Pugan: Physical model-guided underwater image enhancement using gan with dual-discriminators," *IEEE Transactions on Image Processing*, vol. 32, pp. 4472–4485, 2023.
- [61] F. Xiao, J. Liu, Y. Huang, E. Cheng, and F. Yuan, "Semi-supervised underwater image enhancement network boosted by depth map consistency," *IEEE Journal of Oceanic Engineering*, vol. 50, no. 2, pp. 795–816, 2025.

- [62] H. Wang, N. Anantrasirichai, F. Zhang, and D. Bull, "Uwgs: Distractor-aware 3d gaussian splatting for enhanced underwater scene reconstruction," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025, pp. 3280–3289.
- [63] W. Yang, X. Qian, C. H. Lim, Y. Guo, X. Liu, and J. Leng, "Uwgs-slam: A medium-aware underwater visual slam system based on 3d gaussian splatting," *Ocean Engineering*, vol. 356, p. 125143, 2026.
- [64] J. Zhou, Z. He, D. Zhang, S. Liu, X. Fu, and X. Li, "Spatial residual for underwater object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 6, pp. 4996–5013, 2025.
- [65] Z. Chen, H. Gao, Z. Zhang, H. Zhou, X. Wang, and Y. Tian, "Underwater salient object detection by combining 2d and 3d visual features," *Neurocomputing*, vol. 391, pp. 249–259, 2020.
- [66] Q. Li, J. Li, T. Li, Z. Li, and P. Zhang, "Spectral-spatial depth-based framework for hyperspectral underwater target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.
- [67] Y.-T. Peng, Y.-C. Lin, W.-Y. Peng, and C.-Y. Liu, "Blurriness-guided underwater salient object detection and data augmentation," *IEEE Journal of Oceanic Engineering*, vol. 49, no. 3, pp. 1089–1103, 2024.
- [68] Y. Liu, X. Zhang, K. Zhang, B. Ma, S. Yang, R. Yang, and P. Tan, "Detecting underwater salient objects via self-supervised depth priors and task-driven optimization," *Expert Systems with Applications*, vol. 305, p. 130873, 2026.
- [69] J. O. Gaya, L. T. Gonçalves, A. C. Duarte, B. Zanchetta, P. Drews, and S. S. Botelho, "Vision-based obstacle avoidance using deep learning," in *Latin American Robotics Symposium and Brazilian Robotics Symposium*, 2016, pp. 7–12.
- [70] M. Ferrera, V. Creuze, J. Moras, and P. Trouvé-Peloux, "Aqualoc: An underwater dataset for visual-inertial-pressure localization," *International Journal of Robotics Research*, vol. 38, no. 14, pp. 1549–1559, 2019.
- [71] P. Yang, H. Liu, M. Roznere, and A. Q. Li, "Monocular camera and single-beam sonar-based underwater collision-free navigation with domain randomization," in *International Symposium of Robotics Research*. Springer, 2022, pp. 85–101.
- [72] Y. Xie, J. Zhang, N. Bore, and J. Folkesson, "Neurss: Enhancing auv localization and bathymetric mapping with neural rendering for sidescan slam," *IEEE Journal of Oceanic Engineering*, vol. 50, no. 3, pp. 1596–1605, 2025.
- [73] S. Xu, K. Zhang, and S. Wang, "Aqua-slam: Tightly coupled underwater acoustic-visual-inertial slam with sensor calibration," *IEEE Transactions on Robotics*, vol. 41, pp. 2785–2803, 2025.
- [74] Y. Peng, Y. Hong, Z. Hong, A. P.-Y. Chui, and J. Wu, "Aquaticvision: Benchmarking visual slam in underwater environment with events and frames," *arXiv preprint arXiv:2505.03448*, 2025.
- [75] C. Peng, Y. Pan, X. Yang, N. Wang, Z. Yu, and G. Li, "Sa-mpc: Situational awareness-driven model predictive control via vision-sbes fusion for robust underwater obstacle avoidance," *IEEE Journal of Oceanic Engineering*, vol. 51, no. 3, pp. 1928–1941, 2026.
- [76] C. Jiang, H. Zhen, G. Wan, S. Zhu, X. Li, and X. Xiang, "Underwater monocular depth estimation combined with physical prior information," *IFAC Conference on Control Applications in Marine Systems, Robotics and Vehicles*, vol. 59, no. 22, pp. 330–335, 2025.
- [77] A. Phung and R. Camilli, "Sonar-mast3r: Real-time optical-acoustic fusion in turbid, unstructured environments," *arXiv preprint arXiv:2603.13585*, 2026.
- [78] Z. Cai and C. Metzler, "Underwater monocular metric depth estimation: Real-world benchmarks and synthetic fine-tuning with vision foundation models," *arXiv preprint arXiv:2507.02148*, 2025.
- [79] Y. Randall, O. Lifschitz, and T. Treibitz, "Flsea: Underwater visual-inertial and stereovision forward-looking data sets," *Journal of Field Robotics*, 2023.
- [80] N. G. Bach, P. Xuan Tan, and E. Kamioka, "Mutual information-based extrinsic calibration of camera-sonar system leveraging sonar pseudo-pointcloud and underwater light attenuation prior," *IEEE Access*, vol. 14, pp. 40 229–40 249, 2026.