Real-World Benchmarking and Synthetic Fine-Tuning of Monocular Metric Depth Estimation in Underwater Environments

Zijie Cai
University of Maryland, College Park
zai28@umd.edu

Christopher A. Metzler
University of Maryland, College Park
metzler.umd.edu

Abstract

Monocular depth estimation has recently progressed beyond ordinal depth to provide metric depth predictions. However, its reliability in underwater environments remains limited due to light attenuation and scattering, color distortion, turbidity, and the lack of high-quality metric ground-truth data. In this paper, we present a comprehensive benchmark of zero-shot and fine-tuned monocular metric depth estimation models on real-world underwater datasets with metric depth annotations, including FLSea and SQUID. We evaluate a diverse set of state-of-the-art Vision Foundation Models across a range of underwater conditions and depth ranges. Our results show that large-scale models trained on terrestrial data (real or synthetic), while effective in in-air settings, perform poorly underwater due to significant domain shifts. To address this, we fine-tune Depth Anything V2 with a ViT-S backbone encoder on a synthetic underwater variant of the Hypersim dataset, which we simulated using a physically based underwater image formation model. Our fine-tuned model consistently improves performance across all benchmarks and outperforms baselines trained only on the clean in-air Hypersim dataset. This study provides a detailed evaluation and visualization for monocular metric depth estimation in underwater scenes, highlighting the importance of domain adaptation and scale-aware supervision for achieving robust and generalizable metric depth predictions in challenging underwater environments for future research.

1. Introduction

Monocular depth estimation in complex underwater environments is critical for autonomous underwater vehicles (AUVs) for applications such as navigation [37, 68], 3D mapping [45, 64], localization [64, 76], object detection [22, 69], and more. Unlike terrestrial robotics, underwater robotics systems lack dense depth sensing solutions [60]. While LiDARs [15, 39] and RGB-D cameras [20] are widely used above water, their deployment for underwater settings is severely limited by both hardware constraints and high cost [77, 80]. Acoustic Sonar systems are a more common alternative for marine applications [21, 49, 60], but their low spatial resolution due to constrained sensor elevation poses challenges for dense depth perception without additional sensor fusion [32, 48].

Recent advances in monocular depth estimation, particularly those based on vision foundation models using Vision Transformers (ViT) [16] and Dense Prediction Transformers (DPT) [51], have achieved promising performance on in-air datasets like NYUv2 [12] and KITTI [23]. These models typically leverage large-scale real and synthetic RGB-D data to learn powerful scene priors and for both relative and metric depth prediction from a single image [24, 43, 70, 71]. However, their performance degrades significantly in underwater environments due to limited visibility caused by light scattering, turbidity, and wavelength-dependent attenuation [1, 34, 52]. Furthermore, while traditional multi-view methods such as stereo matching [8] or SLAM [17, 61] can recover metric depth of the scene with geometric cues, they require multiple frames and consistent lighting, which are challenging to obtain underwater. In contrast, monocular methods require only a single frame for input, which offers much greater deployment flexibility [41, 72]. However, their zero-shot reliability in underwater settings remains a question [7, 74], not only due to the severe domain shifts in environments [65], but also due to the absence of geometric cues of the scene and the lack of accurate ground-truth metric depth data for supervision [55, 73].

In this work, we aim to address this domain gap by evaluating a diverse set of state-of-the-art monocular metric depth estimation models on two real-world underwater datasets with metric ground truth: FLSea [50] and SQUID [5]. Additionally, we explore the effectiveness of the underwater domain adaptation using a synthetically generated underwater dataset for fine-tuning to improve the model's generalization in underwater performance.

Contributions

Our paper presents a comprehensive benchmark and domain adaptation study of monocular metric depth estimation in underwater environments. Our main contributions are as follows:

- We conduct an extensive zero-shot evaluation of six state-of-the-art monocular metric depth models with varying parameter sizes — including five general-purpose vision foundation models (Depth Anything V2 [71], Metric3D V2 [30], UniDepth V2 [47], ZoeDepth [6], and Depth Pro [7]) and one underwater-specific method (UW-Depth) [18] — on two real-world underwater datasets with metric ground truth: FLSea [50] and SQUID [5].
- We construct a large-scale synthetic underwater dataset by applying a physics-based underwater image formation model [1] to the photorealistic Hypersim RGB-D dataset [53] of simulated indoor scenes with per-pixel ground truth labels, enabling low-cost, in-domain supervision for monocular metric depth estimation.
- We fine-tune Depth Anything V2 (ViT-S) [71] on our synthetically generated underwater version of the Hypersim dataset [53] and demonstrate consistent zero-shot monocular metric estimation improvements across all benchmarks compared to the baseline fine-tuned on the clean in-air Hypersim dataset [53], highlighting the effectiveness of domain-adaptive synthetic training.
- We provide both qualitative and quantitative comparisons across diverse underwater conditions using distinct subset scenes from the FLSea [50] and SQUID [5] datasets, analyzing model robustness and failure cases, and offering insights for future research in underwater perception.

2. Related Work

2.1. Monocular Metric Depth Estimation

Monocular depth estimation has made significant progress in recent years with the increasing availability of large-scale RGB-D datasets [12, 23] and the promising performance of transformer-based vision model architectures [16, 51]. These advancements have extended the development of general-purpose vision foundation models to predict metric depth of a scene in addition to relative depth from a single RGB image [6]. Recent state-of-the-art models such as Depth Anything V2 [71], ZoeDepth [6], and Metric3D V2 [30] all adopt an encoder–decoder architecture [62], where a ViT-based [16] or a convolutional-based [67] encoder extracts high-level scene features and a task-specific decoder predicts dense depth maps [51]. Many of these models are trained using synthetic datasets with ground-truth per-pixel metric depth for supervision, such as Hypersim [53] and SceneNet [38], alongside a large-scale, unlabeled real-world dataset for self-supervised learning with a teacher-student architecture through knowledge distillation [29]. As a result from training, the models demonstrate robust performance for dense depth predictions across both indoor and outdoor in-air scenes by learning strong geometric and semantic priors [71].

2.2. Depth Estimation Benchmarks

Standard benchmarks for monocular depth estimation include NYU Depth v2 (indoor) [12] and KITTI (outdoor) [23], which provide dense metric ground truth for terrestrial scenes. However, few works provide an evaluation of state-of-the-art models in underwater conditions with real metric depth. Our work fills in this gap by evaluating six representative metric depth estimation models (Depth Anything V2 [71], Metric3D V2 [30], UniDepth V2 [47], ZoeDepth [6], Depth Pro [7], UW-Depth [18]) on two real-world underwater datasets (FLSea [50] and SQUID [5]) with consistent qualitative visualizations and quantitative metrics (e.g., AbsRel, δ_1) [19, 28].

2.3. Model Scaling and Inference Trade-offs

Scaling vision models via larger ViT backbones (e.g., ViT-S, ViT-B, ViT-L, ViT-G) generally improves depth prediction accuracy [9, 16, 59, 79]. However, larger models also lead to higher latency in training and inference, and more extensive memory demands, which limit their deployment for real-time applications [4]. Evaluating models across sizes helps identify trade-offs between inference efficiency and accuracy, which is critical for embedded systems such as autonomous underwater vehicles (AUVs), which are often hardware-constrained [72].

2.4. Underwater Depth Estimation

Underwater scenes pose unique challenges for depth estimation due to complex light interactions with water. The visibility in underwater imagery is often limited, caused by issues such as wavelength-dependent attenuation, backscatter, turbidity, and non-uniform illumination [2, 25]. While acoustic sensors such as sonar are commonly used for underwater range sensing, they suffer from low spatial resolution due to limited elevation coverage [48]. Monocular metric depth estimation methods designed specifically for underwater environments, such as UW-Depth [18], are often trained on real-world underwater datasets. However, the scarcity of large-scale, high-quality metric ground-truth data limits model complexity and generalization. As a result, these methods often adopt lightweight encoder–decoder architectures optimized for speed rather than dense accuracy, and their performance tends to degrade outside their training domain [40].

2.5. Synthetic Data for Model Training

To address the lack of ground truth underwater data, recent research has turned to synthetic data generation using

physics-based rendering [3]. The underwater image formation model [1, 25] simulates critical optical effects such as attenuation and backscatter, allowing in-air RGB-D datasets to be transformed into realistic underwater imagery. This approach has been adopted widely in image restoration and enhancement tasks, where obtaining ground-truth color-corrected underwater images is practically infeasible [26]. By allowing supervised training with per-pixel metric depth and controlled variation in water conditions, this synthetic-to-real strategy serves as an effective tool for domain adaptation and model pretraining across various underwater vision tasks [65].

2.6. Domain Adaptation in Vision

Domain adaptation methods aim to improve model generalization across data distributions by aligning features, styles, or learned representations for a new domain [46]. In underwater vision, where the domain shift from terrestrial imagery is severe, prior work has employed strategies such as CycleGAN-based image translation for underwater image enhancement tasks [26, 78]. The authors of *Atlantis: Enabling Underwater Depth Estimation with Stable Diffusion* [73] also propose a pipeline for generating underwater imagery for depth estimation with a diffusion model [13] and a control-net-based approach for training underwater relative depth estimation models [75]. However, such approaches introduce significant overhead in data preparation and add complexity to multi-stage pipelines, which slows down both training and inference for the model.

In contrast, we adopt a forward supervised adaptation strategy [42] by fine-tuning a general-purpose monocular depth model (Depth Anything V2 [71]) on a synthetic underwater dataset generated using a physics-based rendering pipeline to generalize its performance for the underwater metric depth domain. Our supervised fine-tuning strategy (SFT) is lightweight, requiring no auxiliary networks for synthetic data generation, and freezes the early encoder layers to retain the strong pre-trained scene understanding priors while adapting the decoder and late encoder layers to underwater-specific image statistics with labeled data [33].

3. Methods

3.1. Overview

This section outlines our methodology for benchmarking and fine-tuning monocular metric depth estimation in underwater environments. Our pipeline consists of:

1. **Zero-shot evaluation of existing models:** We benchmark a diverse set of six state-of-the-art monocular metric depth estimation models, five of which are general-purpose foundation models and the other one is underwater-specific, on two real-world underwater datasets (FLSea [50] and SQUID [5]).
2. **Synthetic dataset generation:** To address the scarcity of real-world underwater metric ground truth, we create a synthetic underwater dataset by applying a physics-based underwater image formation model to an existing clean in-air synthetic RGB-D dataset (Hypersim [53]). This simulates various underwater imaging conditions while preserving high-quality per-pixel metric depth [1].
3. **Domain adaptation via supervised fine-tuning:** We fine-tune a pretrained ordinal depth foundation model (Depth Anything V2 [71]) using the synthetic underwater Hypersim [53] dataset to adapt the model for predicting underwater metric depth. This approach enables the model to learn underwater-specific visual cues while retaining the generalization ability from pretraining.

3.2. Benchmark Models

We evaluate six state-of-the-art monocular metric depth estimation models, of which five are general-purpose vision foundation models and one is an underwater-specific approach. All models are assessed in a zero-shot setting using publicly available pretrained weights. For in-domain adaptation, we select Depth Anything V2 (ViT-S) [71] as our fine-tuning baseline and explore its performance on the real-world underwater benchmarks.

- **Depth Anything V2 [71]**: A transformer-based depth foundation model originally trained for general-purpose ordinal depth estimation, which can also be fine-tuned for in-domain metric depth estimation. It employs a ViT [16] encoder and DPT [51] decoder and supports multiple encoder scales (ViT-S, ViT-B, ViT-L). We evaluate all three variants that the authors fine-tuned on the clean Hypersim dataset and use ViT-S for our synthetic training.
- **Depth Pro [7]**: A foundation model for zero-shot metric monocular depth and focal length estimation optimized for efficient inference on resource-constrained platforms to provide high-resolution depth maps with unparalleled sharpness and high frequency details.
- **Metric3D V2 [30]**: A geometric foundation model for zero-shot metric depth and surface normal estimation from a single image with the proposed canonical camera space transformation module to address metric ambiguity.
- **UniDepth V2 [47]**: A universal transformer model for zero-shot monocular metric depth estimation that predicts 3D point clouds directly from a single image without requiring camera intrinsics. It employs a self-promptable camera module and geometric invariance losses to predict a dense camera representation for conditioning depth features and enhancing generalization across domains.
- **ZoeDepth [6]**: A scale-aware model that combines ordinal and metric depth cues using a lightweight bin-adjusted head and latent classifier. It is trained on multiple relative and metric datasets to achieve strong zero-shot generalization and scale consistency across domains. Their flag-

ship model, ZoeD-M12-NK, is used for our experiments.
- **UW-Depth [18]**: A lightweight, domain-specific model trained on underwater datasets such as FLSea [50]. It incorporates sparse feature priors to mitigate scale ambiguity for underwater metric depth estimation and is optimized for real-time inference on embedded systems.

3.3. Synthetic Underwater Dataset Generation

To enable supervised fine-tuning in underwater conditions, we generate a synthetic underwater training set based on the Hypersim [53] RGB-D dataset, which provides a dense set of photorealistic indoor scenes and corresponding ground-truth metric depth.

We apply the simplified underwater image formation model for ambient illumination, assuming the camera response is equivalent to the delta function [1]. This formulation accounts for wavelength-dependent attenuation and backscattering. Specifically, we use the wideband approximation of their model, expressed as [10, 27, 56]:

$$I_c = J_c \cdot e^{-\beta_c z} + B_c^{\infty} \cdot (1 - e^{-\beta_c z}), \tag{1}$$

where I_c is the observed underwater intensity in channel $c \in \{R, G, B\}$, J_c is the clear scene radiance, β_c is the wideband attenuation coefficient for channel c, B_c^{∞} is the wideband veiling light (backscatter at infinity), and z is the range (depth) from the camera. This model captures both the exponential attenuation of the direct signal and the accumulation of backscattered light.

We simulate multiple Jerlov water types from open (I, II, III) to coastal ocean classes (1C to 9C) with varying β_c and B_c^{∞} to represent different optical properties and beam absorption levels of ocean water [58]. The resulting dataset consists of paired underwater RGB images and their clean metric depth maps. All images preserve the original pixel alignment and camera intrinsics from Hypersim [53].

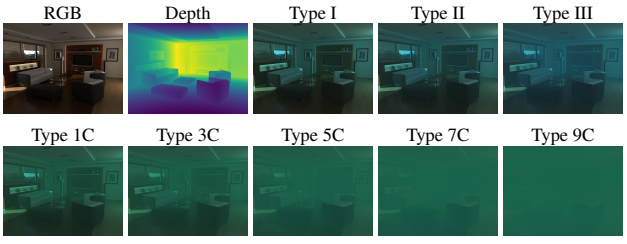

Figure 1. Examples from our synthetic underwater dataset. Top row: clean RGB image, ground-truth depth, and Jerlov open-ocean classes (I, II, III). Bottom row: simulations with increasing attenuation in coastal ocean classes (1C to 9C), representing progressively turbid conditions [58].

3.4. Fine-Tuning Depth Anything on Synthetic Data

We use a supervised fine-tuning strategy using our synthetic underwater dataset in order to adapt the Depth Anything V2 [71] model to the unique underwater imagery domain [65]. This dataset includes varying visual distortions typical of underwater scenes, which each image is one of the Jerlov water classes we saw earlier. Our goal is to enable the model to generalize to the real underwater environment while retaining the strong priors learned from large-scale terrestrial RGB-D and unlabeled data.

We use the ViT-S variant of Depth Anything V2 [71] as our baseline and initialize it with the official relative depth checkpoint. To retain previously learned features, we freeze the first half of the Vision Transformer encoder and allow only the remaining encoder layers and DPT-style decoder to update during training, which is conducted in a supervised regression setting with the following configuration [11, 51]:

- **Optimizer:** AdamW [36] with weight decay of 10^{-2}
- **Learning rate:** 5×10^{-6} with cosine annealing scheduler [35]
- **Warm-up:** Linear warm-up for the first 4 epochs [31]
- **Epochs:** 20 total
- **Batch size:** 4 (constrained by GPU memory)
- **Loss function:** SiLogLoss [19] — a scale-invariant logarithmic loss, commonly used for metric depth regression.
- **Max depth:** 20 meters (used for scaling model output)

We train the model using image–depth pairs at a resolution of 518×518. To enhance robustness for underwater domain adaptation, we apply physically consistent color augmentations, including random illumination changes to simulate under- and overexposure in real-world underwater settings, as well as grayscale conversion to encourage the model to focus on structural cues rather than color variations caused by water conditions [57, 63].

Our supervised fine-tuning strategy enables the model to learn underwater-specific visual cues while retaining general-purpose scene priors from pretraining. By leveraging synthetic data aligned with underwater image formation physics, the fine-tuned model achieves more reliable metric depth estimation in challenging underwater environments.

3.5. Evaluation Datasets

We evaluate all models on two real-world underwater datasets containing RGB images paired with metric ground-truth depth:

- **FLSea [50]:** A large-scale dataset collected in controlled underwater environments using diver-operated cameras and photogrammetry software (Agisoft Metashape) to generate ground-truth metric depth via SFM [44]. We focus on six key subsets from two scenes:
  - *Canyon:* `u_canyon` and `flatiron`, totaling 5,369 images. These subsets capture natural rocky reef structures at water depths of 4–7 meters.
  - *Red Sea:* `big_dice_loop`, `cross_pyramid_loop`, `coral_table_loop`, and `sub_pier`, totaling 6,919 images. These scenes

include both natural structures and large man-made objects such as coral tables, piers, and concrete blocks, with water depths ranging from 3–8 meters.

- **SQUID [5]:** A smaller but more challenging dataset with longer depth ranges and larger areas of featureless background, consisting of 57 stereo image pairs with metric ground-truth depth computed from stereo triangulation [8, 14]. The dataset covers four subset scenes collected across different marine environments in Israel:
  - *Tropical Red Sea:*
    - `Katzaa` — coral reef, 10–15 meters deep.
    - `Satil` — shipwreck site, 20–30 meters deep.
  - *Temperate Mediterranean Sea:*
    - `Nachsolim` — rocky reef, 3–6 meters deep.
    - `Michmoret` — rocky reef, 10–12 meters deep.

All predictions are rescaled to match the dataset-specific depth units. Quantitative evaluations are performed using standard depth estimation metrics, as detailed in Section 4.

4. Experiments

4.1. Evaluation Setup

We evaluate all models on two real-world underwater datasets: FLSea [50] and SQUID [5]. For each model, we use official pre-trained weights and perform zero-shot inference unless otherwise specified. Depth Anything V2 [71] is additionally fine-tuned on our synthetic underwater dataset. When applicable, we evaluate model variants using different ViT encoder sizes (ViT-S, ViT-B, ViT-L).

4.2. Metrics

We report standard monocular depth estimation metrics, widely used in prior work [19, 28]:

- **AbsRel** (Absolute Relative Error): $\frac{1}{|T|}\sum_{i\in T}\frac{|d_i-\hat{d}_i|}{d_i}$
- $\boldsymbol{\delta_1}$: Percentage of predictions satisfying $\delta = \max(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}) < 1.25$

Here, d_i and $\hat{d}_i$ represent the ground-truth and predicted depths, respectively. Metrics are computed only on valid (non-zero) ground-truth pixels and follow the dataset-specific evaluation protocols.

4.3. Zero-Shot Benchmarking Results

We first evaluate all models with official pre-trained weights to compare zero-shot performance to underwater scenes. Table 1 summarizes the results across both datasets.

Notably, several models come with specific limitations in the zero-shot setting. **Metric3D V2** [30] performance varies with input resolution, and we exclude its ViT-L variant due to hardware constraints during inference. **Depth Anything V2** [71], while trained primarily on ordinal depth, provides competitive results but requires further in-domain only fine-tuning to predict metric depth. **UW-Depth** [18] is trained on 10 subsets of the FLSea [50] dataset and two held-out test sets (u_canyon and sub_pier). To avoid evaluation with data leakage, we omit its quantitative results on FLSea [50].

4.4. Effect of Synthetic Fine-Tuning

We fine-tune Depth Anything V2 (ViT-S) [71] on our synthetic underwater dataset and compare its performance to the baseline fine-tuned for in-domain metric depth prediction on the clean Hypersim [53] dataset with the maximum depth scale of 20 meters. This experiment evaluates the effectiveness of forward domain adaptation using synthetic underwater data. Quantitative improvements are summarized in Table 2.

4.5. Qualitative Results

We present qualitative comparisons of predicted depth maps on representative scenes from FLSea [50] and SQUID [5]. Figure 2 showcases zero-shot performance across all benchmarked models using their strongest performing variants, highlighting differences in detail boundary preservation and robustness to underwater artifacts. Figure 3 compares the Depth Anything V2 (ViT-S) [71] baseline with our fine-tuned model, demonstrating the visual impact and effectiveness of synthetic-to-real domain adaptation [65, 66].

5. Discussion

5.1. Zero-Shot Model Performance

Our results demonstrate that general-purpose monocular depth models—such as Depth Anything V2 [71] and UniDepth V2 [47]—perform reasonably well on underwater imagery in a zero-shot setting. However, performance varies significantly across underwater conditions. On datasets with clearer water and narrower depth ranges (e.g., FLSea [50]), models retain moderate accuracy. In contrast, performance degrades in more turbid or visually degraded scenes, such as SQUID [5], with higher *AbsRel* errors and lower δ_1 accuracy [19, 28]. Furthermore, models like ZoeDepth [6] and Depth Pro [7], while robust on benchmarks like NYU-V2 [12] and KITTI [23], experience a significant performance drop when applied to underwater settings. This highlights the difficulty of transferring terrestrial-trained models to underwater domains without explicit adaptation. This is further supported by the performance of the UW-Depth [18] model trained on real-world underwater data with a compact MobileNet V2 backbone [54], yet achieving greater performance than the other larger models. Overall, UniDepth V2 [47] performs exceptionally well on the FLSea [50] dataset, achieving δ_1 accuracy above 90% and the lowest *AbsRel* errors across all benchmarked models [19, 28].

Model	FLSea-Canyon		FLSea-Red Sea		SQUID	
	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
ZoeDepth	1.5907	0.2345	1.3335	0.2109	1.3214	0.0968
Metric3D V2 (ViT-S)†	1.5331	0.0967	0.8130	0.2136	1.3059	0.1680
Depth Pro	0.9858	0.1557	0.3888	0.3772	3.2185	0.1678
UW-Depth†	–	–	–	–	0.4948	0.3446
Depth Anything V2 (ViT-S)†	0.3576	0.4463	0.2569	0.4722	0.5242	0.2054
Depth Anything V2 (ViT-B)†	0.2447	0.5696	0.2471	0.4301	0.4495	0.2649
Depth Anything V2 (ViT-L)†	0.2269	0.6363	0.2307	0.4812	0.3390	0.2896
UniDepth V2 (ViT-S)	0.2233	0.6763	0.1524	0.8122	0.4012	0.4789
UniDepth V2 (ViT-B)	0.1276	0.8844	0.1045	0.9167	0.3638	0.4725
UniDepth V2 (ViT-L)	**0.1156**	**0.9109**	**0.0932**	**0.9439**	**0.3222**	**0.5201**

Table 1. Zero-shot performance comparison across FLSea-Canyon [50], FLSea-Red Sea [50], and SQUID [5] datasets. †Models marked with this symbol are not strictly zero-shot for metric depth. Bold indicates best performance.

Model (ViT-S)	FLSea-Canyon		FLSea-Red Sea		SQUID	
	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
Baseline DA V2	**0.3576**	0.4463	0.2569	0.4722	0.5242	0.2054
Fine-tuned DA V2 (Ours)	0.3620	**0.4683**	**0.2266**	**0.6170**	**0.4465**	**0.3204**

Table 2. Performance of Depth Anything V2 (ViT-S) [71] before and after fine-tuning on synthetic underwater data across FLSea-Canyon [50], FLSea-Red Sea [50], and SQUID [5] subsets. Bold indicates best performance.

5.2. Effectiveness of Synthetic Fine-Tuning

Fine-tuning Depth Anything V2 (ViT-S) [71] on our physics-based synthetic underwater dataset yields consistent improvements across real-world test sets. Quantitatively, fine-tuning reduces absolute errors and improves threshold accuracy on all SQUID [5] scenes and most FLSea [50] subsets. Qualitatively (Fig. 3), the model learns to recover better structural details, suppresses noisy predictions in low-contrast areas, and improves depth continuity in visually degraded regions. This underscores the utility of synthetic domain adaptation for underwater vision tasks where collecting metric-labeled real data is costly or infeasible [26, 52, 65].

5.3. Qualitative Observations

Across zero-shot and fine-tuned visual depth map comparisons (Fig. 2 and Fig. 3), we observe:

- **UniDepth V2** [47]: Among all evaluated models, UniDepth V2 [47] consistently delivers the best overall performance across all benchmarks. It produces accurate metric depth maps with fine structural details and strong generalization ability to varying underwater conditions.
- **Depth Anything V2 (fine-tuned)** [71]: The baseline Depth Anything V2 [71] ranks just behind UniDepth V2 [47], showing strong metric scale accuracy but slightly less precise boundary detail. After fine-tuning the ViT-S variant on our synthetic underwater dataset, the model exhibits improved structural consistency, sharper depth boundaries, and more accurate metric depth predictions.
- **UW-Depth** [18]: While the compact backbone design limits spatial resolution and leads to loss of structural detail in predicted depth maps, the model still provides reasonably accurate metric scale predictions, especially in scenes similar to its training distribution.
- **Metric3D V2** [30]: The ViT-L variant of this model visually performs on par with UniDepth. However, it is difficult to confirm its exact performance without any quantitative evaluation, highlighting one of our key limitations.
- **ZoeDepth [6] and Depth Pro [7]**: While producing relatively smooth depth maps with decent edge boundary preservation, these models tend to have more inaccuracies in metric scale prediction compared to others.

5.4. Model Trade-offs

Transformer-based models with larger encoders (e.g., ViT-L) typically offer better zero-shot accuracy and generalization. However, they also introduce extra computational overhead, which can make them less practical for real-time deployment on embedded systems [54]. Lightweight models such as UW-Depth [18] provide a viable speed-accuracy trade-off, which suggests a gap between compact model efficiency and underwater robustness.

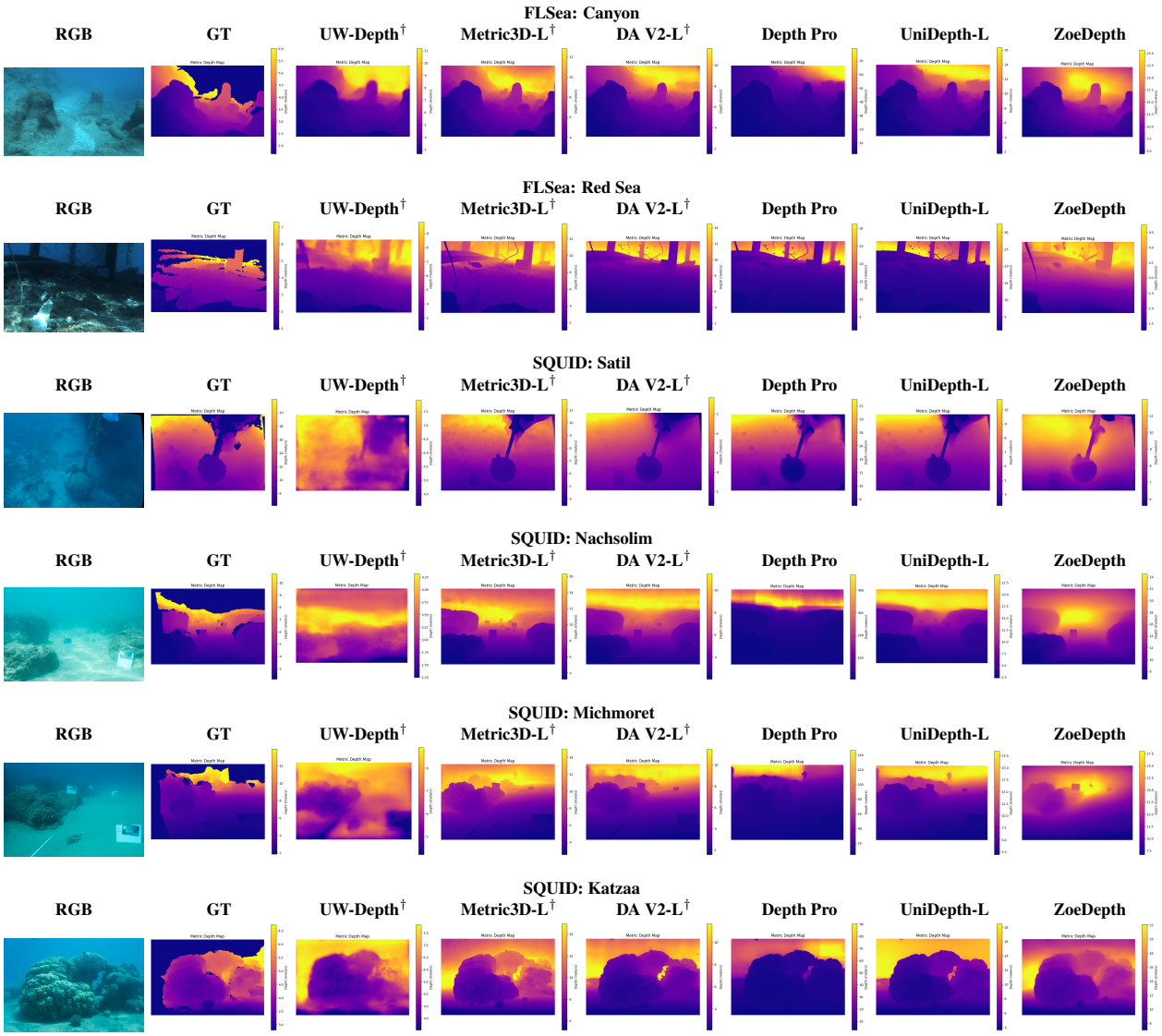

Figure 2. Qualitative comparisons on 6 underwater scenes from two real-world underwater datasets: FLSea and SQUID. Each group shows the RGB input, ground-truth (GT), UW-Depth† [18], Metric3D V2† [30], Depth Anything V2 (ViT-L)† [71], Depth Pro [7], UniDepth V2 (ViT-L) [47], and ZoeDepth [6]. UniDepth V2 (ViT-L) [47] consistently produces the most accurate metric depth maps with fine edge details across all datasets. Depth Anything V2 (ViT-L) [71] also performs competitively, particularly in near-range scenes (<20 meters), but shows softer boundaries and more artifacts at longer ranges. Metric3D V2 (ViT-L) [30] outputs visually sharpest maps, but its metric depth scale accuracy is unclear due to the lack of its quantitative results in our study. Depth Pro [7] is affected the most by texture-less regions in the background, leading to poor structural recovery for the foreground. ZoeDepth [6] and UW-Depth [18] both underperform due to their limited model capacity and training domain scope.

5.5. Limitations and Future Work

While synthetic fine-tuning notably improves performance, several challenges remain:

- Performance significantly degrades in extreme conditions, such as highly turbid water, low-light scenes, or regions with texture-less backgrounds.
- We only fine-tuned Depth Anything V2 (ViT-S) [71]; other backbone variants and models (e.g., UniDepth [47]) could also benefit from synthetic adaptation.
- Our synthetic dataset is limited to Hypersim-based indoor geometry [53]; future work could incorporate more diverse 3D structures and scene ranges. Also, simulating additional real-world underwater phenomena with stable diffusion and using unlabeled underwater images for self-supervised training could enhance model generalization for the underwater setting as well [73].

Overall, our benchmark highlights both the promise and limitations of monocular depth estimation in underwater conditions and provides understanding for further exploration in data simulation, cross-model adaptation, and real-world deployment in future research.

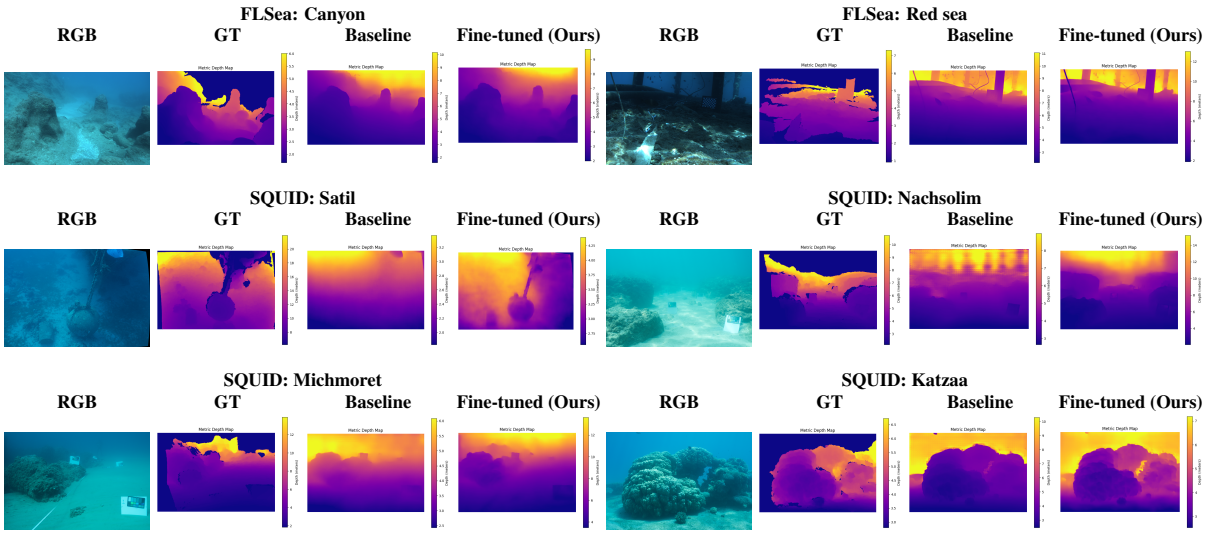

Figure 3. Qualitative comparison between the baseline (Depth Anything V2 ViT-S [71]) and our fine-tuned model using synthetic underwater data on six scenes from two real-world datasets: FLSea [50] and SQUID [5]. Each group shows the RGB input, ground-truth (GT), zero-shot baseline prediction, and prediction after fine-tuning. Our synthetically fine-tuned models produce much sharper depth boundaries and a more accurate metric depth scale with improved robustness for adapting underwater domain than the baseline models across all scenes, especially in turbid and low-contrast regions seen in the SQUID[5] dataset with high scattering and color distortion.

6. Conclusion

In this work, we presented a comprehensive benchmark of monocular metric depth estimation models for underwater environments, comparing a diverse set of state-of-the-art general-purpose and domain-specific approaches across two challenging real-world datasets: FLSea [50] and SQUID [5], each representing distinct underwater conditions in terms of visibility, depth ranges, and scene complexity. Our goal was to evaluate the models' zero-shot generalization capability and explore whether physics-based synthetic data can effectively enable underwater domain adaptation.

We developed a synthetic data generation pipeline that simulates realistic underwater RGB images from Hypersim [53] using a physics-based underwater image formation model [1] to address the lack of large-scale, annotated high-quality real-world underwater datasets. This pipeline incorporates varying wavelength-dependent attenuation and backscattering, generating paired RGB-depth data across multiple Jerlov water types [58].

Our results reveal that while general-purpose models (e.g., UniDepth V2 [47], Depth Anything V2 [71], Metric3D V2 [30], ZoeDepth [6]) show moderate zero-shot performance in clear water scenes with narrow depth ranges and textural information, their accuracy drops significantly in visually degraded or turbid conditions. Among all models evaluated, UniDepth V2 [47] achieves the best zero-shot performance across all datasets, particularly in preserving metric scale and structural consistency.

We further demonstrate that fine-tuning Depth Anything V2 (ViT-S) [71] on our synthetic underwater dataset improves both quantitative and qualitative performance, especially in low-visibility scenarios, where the baseline fails to extract structural details. The fine-tuned model produces sharper boundaries, better depth consistency, and more accurate metric predictions, confirming the effectiveness of using synthetic data for underwater domain adaptation.

Overall, our benchmark reveals the difficulty of transferring general-purpose depth models trained with mainly terrestrial data to underwater settings and the potential of synthetic data to bridge this domain gap. Future directions include: (1) extending fine-tuning to additional backbone variants (e.g., ViT-L, MobileNetV2, etc) and models like UniDepth [47], (2) incorporating more diverse and dynamic underwater scenes, and (3) more detailed evaluation of inference speed and accuracy trade-off to help select the appropriate model on resource-constrained platforms for real-time deployments. All of these efforts will be critical to enhance the robustness of monocular metric depth models in practical real-world underwater applications.

Acknowledgments

I would like to thank Dr. Christopher Metzler for his guidance and feedback throughout this project, and Tianfu Wang and members of the Intelligent Sensing Laboratory at the University of Maryland, College Park, for helpful discussions and suggestions during development.

References

[1] Derya Akkaynak and Tali Treibitz. A revised underwater image formation model. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6723–6732, 2018. 1, 2, 3, 4, 8

[2] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[3] Abdelhakim Amer, Olaya Álvarez Tuñón, Halil İbrahim Uğurlu, Jonas Le Fevre Sejersen, Yury Brodskiy, and Erdal Kayacan. Unav-sim: A visually realistic underwater robotics simulator and synthetic data-generation framework. In *2023 21st International Conference on Advanced Robotics (ICAR)*, pages 570–576, 2023. 3

[4] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2245–2264, 2025. 2

[5] Dana Berman, Deborah Levy, Shai Avidan, and Tali Treibitz. Underwater single image color restoration using haze-lines and a new quantitative dataset, 2018. 1, 2, 3, 5, 6, 8

[6] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 2, 3, 5, 6, 7, 8

[7] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second, 2025. 1, 2, 3, 5, 6, 7

[8] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 1, 5

[9] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021. 2

[10] John Y. Chiang and Ying-Ching Chen. Underwater image enhancement by wavelength compensation and dehazing. *IEEE Transactions on Image Processing*, 21(4):1756–1769, 2012. 4

[11] Brian Chu, Vashisht Madhavan, Oscar Beijbom, Judy Hoffman, and Trevor Darrell. Best practices for fine-tuning visual classifiers to new domains. In *Computer Vision – ECCV 2016 Workshops*, pages 435–442, Cham, 2016. Springer International Publishing. 4

[12] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013. 1, 2, 5

[13] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023. 3

[14] J. Davis, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: a unifying framework for depth from triangulation. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages II–359, 2003. 5

[15] Pinliang Dong and Qi Chen. *LiDAR remote sensing and applications*. CRC Press, 2017. 1

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 3

[17] Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110, 2006. 1

[18] Luca Ebner, Gideon Billings, and Stefan Williams. Metrically scaled monocular depth estimation through sparse priors for underwater robots, 2023. 2, 4, 5, 6, 7

[19] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. 2, 4, 5

[20] Felix Endres, Jürgen Hess, Jürgen Sturm, Daniel Cremers, and Wolfram Burgard. 3-d mapping with an rgb-d camera. *IEEE transactions on robotics*, 30(1):177–187, 2013. 1

[21] Maurice F Fallon, John Folkesson, Hunter McClelland, and John J Leonard. Relocating underwater features autonomously using sonar-based slam. *IEEE Journal of Oceanic Engineering*, 38(3):500–513, 2013. 1

[22] Sheezan Fayaz, Shabir A Parah, and GJ Qureshi. Underwater object detection: architectures and algorithms–a comprehensive review. *Multimedia Tools and Applications*, 81(15): 20871–20916, 2022. 1

[23] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013. 1, 2, 5

[24] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 1

[25] Salma P. González-Sabbagh and Antonio Robles-Kelly. A survey on underwater computer vision. *ACM Comput. Surv.*, 55(13s), 2023. 2, 3

[26] Praful Hambarde, Subrahmanyam Murala, and Abhinav Dhall. Uw-gan: Single-image depth estimation and image enhancement for underwater images. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021. 3, 6

[27] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2011. 4

[28] Xiankang He, Dongyan Guo, Hongji Li, Ruibo Li, Ying Cui, and Chi Zhang. Distill any depth: Distillation creates a stronger monocular depth estimator, 2025. 2, 5

[29] Chengming Hu, Xuan Li, Dan Liu, Haolun Wu, Xi Chen, Ju Wang, and Xue Liu. Teacher-student architecture for knowledge distillation: A survey, 2023. 2

[30] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024. 2, 3, 5, 6, 7, 8

[31] Dayal Singh Kalra and Maissam Barkeshli. Why warmup the learning rate? underlying mechanisms and improvements, 2024. 4

[32] Hong-Gi Kim, Jungmin Seo, and Soo Mee Kim. Underwater optical-sonar image fusion systems. *Sensors*, 22(21):8445, 2022. 1

[33] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution, 2022. 3

[34] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE transactions on image processing*, 29:4376–4389, 2019. 1

[35] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. 4

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 4

[37] Yuncheng Lu, Zhucun Xue, Gui-Song Xia, and Liangpei Zhang. A survey on vision-based uav navigation. *Geospatial information science*, 21(1):21–32, 2018. 1

[38] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J. Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[39] Dan McLeod, John Jacobson, Mark Hardy, and Carl Embry. Autonomous inspection using an underwater 3d lidar. In *2013 OCEANS-San Diego*, pages 1–8. IEEE, 2013. 1

[40] Sachin Mehta and Mohammad Rastegari. Mobilevit: Lightweight, general-purpose, and mobile-friendly vision transformer, 2022. 2

[41] Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33, 2021. 1

[42] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[43] Sergey I Nikolenko et al. *Synthetic data for deep learning*. Springer, 2021. 1

[44] Onur Ozyesil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion, 2017. 4

[45] Narcís Palomeras, Natalia Hurtós, Marc Carreras, and Pere Ridao. Autonomous mapping of underwater 3-d structures: From view planning to execution. *IEEE Robotics and Automation Letters*, 3(3):1965–1971, 2018. 1

[46] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015. 3

[47] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler, 2025. 2, 3, 5, 6, 7, 8

[48] Ziyuan Qu, Omkar Vengurlekar, Mohamad Qadri, Kevin Zhang, Michael Kaess, Christopher Metzler, Suren Jayasuriya, and Adithya Pediredla. Z-splat: Z-axis gaussian splatting for camera-sonar fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2

[49] Sharmin Rahman, Alberto Quattrini Li, and Ioannis Rekleitis. Svin2: An underwater slam system using sonar, visual, inertial, and depth sensor. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1861–1868. IEEE, 2019. 1

[50] Yelena Randall and Tali Treibitz. *FLSea: Underwater Visual-Inertial and Stereo-Vision Forward-Looking Datasets*. PhD thesis, ProQuest Dissertations Publishing, 2023. PQDT - Global, ISBN: 9798379478926. 1, 2, 3, 4, 5, 6, 8

[51] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 1, 2, 3, 4

[52] Smitha Raveendran, Mukesh D Patil, and Gajanan K Birajdar. Underwater image enhancement: a comprehensive review, recent trends, challenges and applications. *Artificial Intelligence Review*, 54:5413–5467, 2021. 1, 6

[53] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021. 2, 3, 4, 5, 7, 8

[54] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019. 5, 6

[55] Ashutosh Saxena, Jamie Schulte, and Andrew Y. Ng. Depth estimation using monocular and stereo cues. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, page 2197–2203, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc. 1

[56] Y.Y. Schechner and N. Karpel. Clear underwater vision. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, pages I–I, 2004. 4

[57] Yangming Shi, Binquan Wang, Xiaopo Wu, and Ming Zhu. Unsupervised low-light image enhancement by extracting structural similarity and color consistency. *IEEE Signal Processing Letters*, 29:997–1001, 2022. 4

[58] Michael G. Solonenko and Curtis D. Mobley. Inherent optical properties of jerlov water types. *Appl. Opt.*, 54(17): 5392–5401, 2015. 4, 8

[59] Andreas Steiner, Alexander Kolesnikov, , Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 2

[60] Kai Sun, Weicheng Cui, and Chi Chen. Review of underwater sensing technologies and applications. *Sensors*, 21(23): 7849, 2021. 1

[61] Takafumi Taketomi, Hideaki Uchiyama, and Sei Ikeda. Visual slam algorithms: A survey from 2010 to 2016. *IPSJ transactions on computer vision and applications*, 9(1):16, 2017. 1

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 2

[63] Sophia J. Wagner, Nadieh Khalili, Raghav Sharma, Melanie Boxberg, Carsten Marr, Walter de Back, and Tingying Peng. Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 257–266, Cham, 2021. Springer International Publishing. 4

[64] Jinkun Wang, Shi Bai, and Brendan Englot. Underwater localization and 3d mapping of submerged structures with a single-beam scanning sonar. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4898–4905. IEEE, 2017. 1

[65] Zhengyong Wang, Liquan Shen, Mai Xu, Mei Yu, Kun Wang, and Yufei Lin. Domain adaptation for underwater image enhancement. *IEEE Transactions on Image Processing*, 32:1442–1457, 2023. 1, 3, 4, 5, 6

[66] Junjie Wen, Jinqiang Cui, Zhenjun Zhao, Ruixin Yan, Zhi Gao, Lihua Dou, and Ben M. Chen. Syreanet: A physically guided underwater image enhancement framework integrating synthetic and real images. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5177–5183, 2023. 5

[67] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16133–16142, 2023. 2

[68] Marios Xanthidis, Nare Karapetyan, Hunter Damron, Sharmin Rahman, James Johnson, Allison O'Connell, Jason M. O'Kane, and Ioannis Rekleitis. Navigation in the presence of obstacles for an agile autonomous underwater vehicle. *arXiv preprint arXiv:1903.11750*, 2020. 1

[69] Shubo Xu, Minghua Zhang, Wei Song, Haibin Mei, Qi He, and Antonio Liotta. A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing*, 527:204–232, 2023. 1

[70] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 1

[71] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[72] Boxiao Yu, Jiayi Wu, and Md Jahidul Islam. Udepth: Fast monocular depth estimation for visually-guided underwater robots. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3116–3123, 2023. 1, 2

[73] Fan Zhang, Shaodi You, Yu Li, and Ying Fu. Atlantis: Enabling underwater depth estimation with stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11852–11861, 2024. 1, 3, 7

[74] Jiuling Zhang. Survey on monocular metric depth estimation, 2025. 1

[75] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3

[76] Pengfei Zhang, Zhengxing Wu, Jian Wang, Shihan Kong, Min Tan, and Junzhi Yu. An open-source, fiducial-based, underwater stereo visual-inertial localization method with refraction correction. In *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4331–4336. IEEE, 2021. 1

[77] Guoqing Zhou, Chenyang Li, Dianjun Zhang, Dequan Liu, Xiang Zhou, and Jie Zhan. Overview of underwater transmission characteristics of oceanic lidar. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:8144–8159, 2021. 1

[78] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. 3

[79] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. *ICLR*, 2022. 2

[80] SF Zhuang, Y Ji, DW Tu, and X Zhang. Underwater rgb-d camera based on binocular stereo vision. *Acta Photonica Sin*, 51:0404003, 2022. 1