Scaling In-Air Monocular Depth Models for Underwater Environments Using Single-Point Ground Truth

Shuzheng Zhang
shz078@ucsd.edu
University of California, San Diego
La Jolla, California, USA

Christopher L. Crutchfield
ccrutchf@ucsd.edu
University of California, San Diego
La Jolla, California, USA

Sijan Shrestha
sijan.shrestha@bison.howard.edu
Howard University
Washington, District of Columbia, USA

Curt Schurgers
cschurgers@ucsd.edu
University of California, San Diego
La Jolla, California, USA

Ryan Kastner
kastner@ucsd.edu
University of California, San Diego
La Jolla, California, USA

Abstract

Reliable underwater depth maps are important for many underwater tasks, but commercial sensors and terrestrial monocular depth models often fail because of backscattering, light loss, and water turbidity. Modern in-air depth estimation models like Depth Anything 3 and Depth Pro retain strong relative scene depth information when processing underwater images, and accurate metric scale can be recovered without the high computational cost of full training. This study evaluates recovery accuracy by scaling the depth map to known ground truth and introduces a Multilayer Perceptron (MLP) scale predictor with a ConvNeXt backbone, using single-point ground truth to predict a scaling factor and recover metric depth with low training compute cost. Evaluations on the SeaThru dataset shows significant accuracy improvement over unscaled outputs, and good metric results when depth maps are scaled to a single-point ground truth.

CCS Concepts

• **Applied computing** → **Environmental sciences**; Systems biology; *Physics*; • **Computing methodologies** → **Neural networks**; *Reconstruction*.

Keywords

Underwater Monocular Depth, Monocular Depth Estimation

ACM Reference Format:
Shuzheng Zhang, Christopher L. Crutchfield, Sijan Shrestha, Curt Schurgers, and Ryan Kastner. 2026. Scaling In-Air Monocular Depth Models for Underwater Environments Using Single-Point Ground Truth. In *Proceedings of International Conference on Underwater Networks and Systems (WUWNet'26).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WUWNet'26, Hong Kong, China
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2026/12
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Accurate underwater depth measurement is important for a wide range of marine applications, including fisheries stock assessment where length-frequency distributions are crucial to estimate ecological health [11], benthic habitat monitoring where highly detailed surface information is essential to quantify habitat structural complexity [10], and digital reconstruction of reefs to better understand the coral ecosystem [5]. Autonomous underwater robotics also require depth map to make operation decisions such as obstacle avoidance and object recognition, and to collect data for tasks such as seafloor mapping and infrastructure inspection.

Methods commonly used to measure accurate depth maps in air fail in water since water aggressively absorbs the near-infrared light emitted by active sensors [16]. Additionally, water turbidity and clarity, backscatter, and noise reduce the range and effectiveness of underwater LiDAR sensors and commercial stereo cameras, producing highly inaccurate results and sometimes failing to calibrate fully [16, 18]. Specialized stereo camera setups are commonly used, but they require two calibrated waterproof cameras, making the collection of high-quality underwater depth maps at scale costly. The FishSense system [7] enables citizen scientists to tackle this problem by using a consumer camera paired with a laser to collect depth measurements, but only one point of depth is measured.

There exist many good, low-cost technologies for in-air depth maps, most promisingly monocular depth vision transformers, with Depth Anything 3 (DA3) [12] and Depth Pro [4] among the most accurate. However, these terrestrial models degrade under the same effects, leading to poor metric depth estimates [6]. Previous studies show that although their unprocessed estimates are poor, zero-shot monocular depth models trained on in-air photos retain good relative depth information when processing underwater images, and high-accuracy metric depth maps can be retrieved if some prior ground truth is known [6, 21]. This study focuses on DA3 and Depth Pro, investigating methods to use single-point depth from a potentially commercially viable camera to train and predict a scaling factor that recovers metric depth from their output.

Our primary contribution is a computationally efficient method to adapt terrestrial monocular depth models for underwater images using single-point depth ground truth, both as a direct scale reference and by training a Multilayer Perceptron (MLP) scale predictor.

We treat median scaling over the full ground truth only as an upper bound, and evaluate the predictor on held-out frames against unmodified baselines. The predictor improves the estimation, but fails to reproduce results we deem accurate enough for scientific uses. However, single-point scaling significantly improves the metric depth accuracy of in-air models underwater, and could have implications in the future as these models improve.

2 Related Work

Significant advancements have been made in underwater monocular depth estimation. Physics-based approaches such as UPGformer integrate light interactions underwater to handle the light and color loss caused by water to achieve good metric depth estimations [17]. Geometry-based approaches trained using camera motion and triangulation have also shown promising results [2, 8], and most accurate underwater depth map datasets are derived from photogrammetry or calibrated stereo cameras [1, 3, 15].

However, because underwater conditions are relatively volatile, models trained specifically on underwater data must be retrained for different water conditions, making both the data and the models costly to produce, both monetarily and computationally. To address this issue, recent methods use synthetically generated images for training to offset data-collection costs. For example, the Atlantis [20] framework generates underwater scenes with Stable Diffusion and ControlNet; however, due to its lack of physical calculations, accuracy cannot be guaranteed at scale.

Furthermore, these challenges cause underwater monocular depth estimation to lag substantially behind its terrestrial counterpart [20], and adapting in-air models can produce accurate metric depth results for underwater scenes at low compute cost. PUDE [19] transfers in-air depth to underwater scenes with unknown depth using a physical image formation model that aligns with water properties. Yang et al.'s method achieves good accuracy using only 76 images in three epochs; however, it outputs relative depth, which requires ground truth maps during evaluation, making it unsuitable for low-cost systems.

3 Methods

The primary objective of this paper is to investigate if modern terrestrial monocular depth models can be efficiently adapted for underwater environments using single-point depth data produced from a low-cost system. We study two deployment cases where a single-point depth is available for every image, and where it is no longer available to simulate a deployment of such a device. We perform experiments on the SeaThru dataset with ground truth derived from underwater Structure-from-Motion [1] and assess the accuracy of in-air monocular depth models after scaling, specifically DA3's DA3NESTED-GIANT-LARGE-1.1 model [12] and Depth Pro's standard model [4].

To train a predictor when depth information is no longer available, we train MLP blocks and attach them to the end of a ConvNeXt-Large model [13], replacing its image classification head, where it functions as a scaling predictor for the monocular depth models. This predicted scale is applied to recover the metric depth map underwater. We then evaluate the accuracy of the output with the ground truth and assess their performance. This method was chosen because of its low compute cost and potential uses on data produced by the FishSense Lite system [7]. We freeze the large monocular depth models and most of the ConvNeXt network, requiring only one point of depth data to train, which also lets us apply it to our existing data in the future.

3.1 Evaluation

We evaluate model accuracy by comparing its output against the ground truth using standard Root Mean Squared Error (*RMSE*), Absolute Relative error (*AbsRel*), the threshold accuracy metric (δ_1) [9], and the Mean Absolute Error (*MAE*). We evaluate using the equations as follows, where N is the subset of valid pixels in the depth maps:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{d}_i - d_i^*)^2} \qquad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{d}_i - d_i^*|$$

$$\text{AbsRel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{d}_i - d_i^*|}{d_i^*}$$

$$\delta_1 = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\max\left(\frac{\hat{d}_i}{d_i^*}, \frac{d_i^*}{\hat{d}_i}\right) < 1.25\right)$$

RMSE penalizes larger errors more heavily, whereas MAE reflects average errors linearly, providing a contrast to RMSE. AbsRel provides a distance-normalized value on the error, and the threshold accuracy (δ_1) reflects the percentage of pixels predicted within 25% of the ground truth. These values are calculated for every image in a dataset; we then take their mean to evaluate overall accuracy.

3.2 Scaling Accuracy

We improve these results by treating the metric depth outputs as relative depth and applying some scaling factor s. For each image, s is the ratio between the median of the ground truth and the median of the output depth map:

$$s = \frac{\text{median}(d^*)}{\text{median}(\hat{d})}$$

where d^* is the ground truth depth from the SeaThru dataset, and $\hat{d}$ is the predicted depth from the monocular depth model output. We will then apply it to the entire depth map from the monocular depth model output:

$$d_{scaled} = s \times \hat{d}$$

We will then treat the resulting d_{scaled} as the $\hat{d}$ of the respective method when evaluating its accuracy. Because this median uses the full ground-truth map, we treat it as an oracle, an upper bound on scale-only correction.

3.2.1 Unmodified Results. Table 1 shows poor unmodified accuracy from DA3 and Depth Pro. Dataset 5 is an exception, where DA3 performs notably better than the other datasets, partially due to roughly half the dataset being far and unmeasurable, resulting in NaN for these values. We also noticed both models struggle most with Datasets 1 and 3, which are taken in low backscattering (B_c) environments [1].

		Unmodified				Scaled			
	Set	RMSE↓	MAE↓	AbsRel↓	δ_1↑	RMSE↓	MAE↓	AbsRel↓	δ_1↑
DA3	D1	1.778	1.762	1.782	0.024	0.203	0.156	0.150	0.801
	D2	2.363	2.266	0.454	0.207	0.613	0.504	0.172	0.994
	D3	3.557	2.954	2.176	0.043	0.807	0.292	0.187	0.910
	D4	1.692	1.656	0.905	0.167	0.304	0.231	0.279	0.941
	D5	2.714	1.557	3.579	0.608	2.494	1.109	3.708	0.850
DPro	D1	2.159	2.125	1.985	0.056	0.169	0.129	0.127	0.873
	D2	1.653	1.563	0.361	0.375	0.458	0.369	0.152	0.981
	D3	0.902	0.510	0.381	0.477	0.819	0.295	0.237	0.909
	D4	0.958	0.914	0.489	0.363	0.253	0.189	0.277	0.970
	D5	3.706	2.136	4.009	0.395	3.781	1.252	4.682	0.873

Table 1: Unmodified DA3 and Depth Pro (DPro) results across all SeaThru Datasets, and their median scaled results respectively

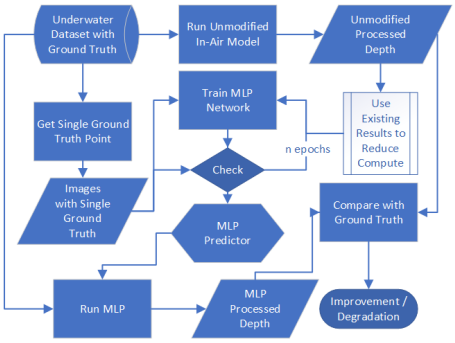


Figure 1: Training and Applying the MLP Predictor

After calculating the scaling factor s from the medians of each depth map and applying it, we saw a substantial increase in both models on most datasets, except Dataset 5, which saw a lower percentage increase but a good accuracy value. Percentage improvements are included in [22]. Overall, we saw minor improvements and slight deterioration in AbsRel after scaling in both models. We observed over a 100% improvement in most threshold-accuracy (δ_1) metrics and over 1000% improvement in some datasets after scaling. We also saw a greater than 60% improvement in RMSE, MAE, and AbsRel for most datasets, resulting in values below 1 for all but the aforementioned Dataset 5 in these metrics, and more than 0.85 for almost all datasets in δ_1. A positive improvement indicates lower RMSE, MAE, and AbsRel, but higher δ_1.

3.2.2 Median Scaled Results. These in-air models retain a strong relative depth underwater in the SeaThru Dataset, as the scene information remains intact and, in most cases, we can recover substantial metric depth data by applying a scaling factor s to the unmodified metric depth output. This occurs because relative geometric features remain intact for the neural network; however, complex localized distortions introduces inaccuracies this method cannot resolve.

3.3 Scale Predictor

We propose a MLP block trained with single-point depths to predict scale s for an underwater image under similar water conditions.

3.3.1 Single-Point Ground Truth. To mimic a single-point depth camera similar to the FishSense Lite system [7], we generate a CSV that provides exactly one coordinate for each image and its respective ground truth from the SeaThru depth map. We locate this pixel by setting the image center as the origin and searching the closest valid depth within 150 pixels above and/or left of the origin [1], then retrieve the depth from the ground truth; if no valid pixel is found, this image is skipped.

3.3.2 Training Process. To reduce compute, we use the existing depth maps produced by the models and the table of single-point ground truth for each dataset, as well as their respective image as input. We resize the images to a set dimension for ConvNeXt and partition them into 85% training and 15% validation, before entering the iterative training loop, shown in Figure 1.

For every epoch, each batch is processed through a ConvNeXt backbone and an MLP head to predict a scalar value; we then take the median of a 5 x 5 pixel patch from the unprocessed output as rescaling the image resolution causes slight inaccuracies to pixel location, and generate the metric depth scale prediction. We compute an $L1$ Loss between these derived metric predictions and the ground-truth depths, which is used to update the model parameters using the AdamW optimizer [14] with a learning rate of 10^{-4}. We then check the model and calculate the validation loss in MAE; if it reaches a new minimum, the new weights are saved.

Most training and inference were done on workstation-class GPUs, timing and device specifics can be found at [22].

4 Results

We separate the results into single-point scaling results and scale-predictor scaling results. The single-point results emulate the scenario similar to FishSense Lite [7] where a single ground truth is available, and the scale predictor results represent the case where this information is not available, and an MLP predictor was used. Dataset n is abbreviated as Dn below, and SDn in supporting data [22].

4.1 Single-Point Results

We replaced the median scale by scaling with this single-point ground truth against the predicted depth by the monocular depth models, and ran the evaluation of the models again to obtain the accuracy data seen in [22]. In general, as shown in Figure 2, we found similar levels of improvement with the median scaling results, with good results in all datasets except D5. This is caused by the photo being taken parallel to the ocean floor, which means the emulated laser dot lands in places with depth that varies greatly, even when the pixel coordinates are relatively close: minimum of 2.17m, Q1 of 7m, Q3 of 10.41m, and max of 17.76m. This scenario is impossible for a diver-operated laser system, so we disregard this result.

Therefore, we conclude that scaling predicted depth by matching single-point ground truth effectively increases the performance of in-air metric monocular depth models underwater.

[1] This is chosen as it is unrealistic for the laser to be further than 150px away from center by parallax; see [7] for system geometry

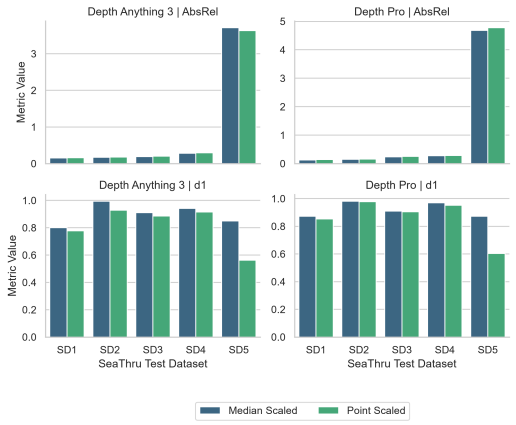


Figure 2: Median-Scaled compared to Point-Scaled Accuracy Metrics Across All SeaThru Datasets

4.2 Scale Predictor Results

We trained separate MLP predictors for SeaThru D1 through D3, as well as a subset of the first 100 images for D1 and D2, and the first 50 images for the smaller D4[2] to simulate the model being used on newly taken photos; we defined these as Dn_{sub_x}. We also created a mixed dataset DX containing the first 50 images from each dataset. Each predictor was run on all SeaThru Datasets to evaluate general performance. We also trained model on roughly uniformly sampled subsets $Dn_{scatter_x}$ with all data and associated benchmarks at [22].

All weights significantly improved whole-image depth accuracy over the exact training datasets, as the data provide some known prior for the exact images for relative depth scaling. Although not as significant, we also found large uplifts with most Dn_{sub_x} models on their respective Dn datasets; this is because the water caused a similar amount of distortion in these similar locations with almost identical water conditions. The exception is D2 in Depth Pro, which showed a minor degradation. The improvements over unseen photos are also strong for D1 and D4, but not for D2. We investigate these in Section 4.2.1.

Although the uplifts are significant, the accuracy itself is mostly fair, with some outlying good results. Currently, there is no paper linking underwater monocular depth accuracy directly to the acceptable error range for fish population health surveying; however, with a best case accuracy of 0.363 RMSE, 0.320 MAE, 0.331 AbsRel, and 0.487 δ_1 seen in the $D1_{sub_{100}}$ model with DA3, it is unsuitable for this role. Therefore, although we saw significant uplifts of up to around 2000%, we cannot conclude that training such a scale predictor can produce good scientific estimates from in-air monocular depth results for underwater images in similar locations. We expect more significant uplifts with a dataset where the camera is stationary in one exact location throughout all photos, as it is difficult to isolate the impact of camera movement on accuracy, but unfortunately no dataset of such with highly accurate depth maps exists today due to the restrictions with underwater depth mapping.

[2]D3 and D5 were too small to train on a subset of images

We found a loose correlation between the improvement in accuracy metrics and the similarity between training and inference dataset water conditions. For DA3 only, we found substantial uplift in all low backscatter datasets with all weights trained on low backscatter data; however, this improvement is not consistent with weights trained on high backscatter data. This may be caused by the two high backscatter datasets being taken on different cameras with different optical characteristics [1], but we do not have sufficient data to confirm this. We also cannot find the same pattern with results from Depth Pro.

All other results were mixed with mostly degradations in accuracy, notably when the backscatter in the data specific MLP is trained on does not match the dataset being run on, where some δ_1 reach zero as no pixels on average are near 25% of each other. This is expected, as we found photos with low backscatter in this dataset typically require a fractional scaling factor less than 1: $s_{low} = \frac{1}{n}$, and the inverse $s_{high} = n, \{n \in \mathbb{R}^+ | n \geq 1\}$. Therefore, a mismatch in water conditions would cause a degradation in accuracy.

4.2.1 Scattered Training Set. We investigated the exception in D2 and found that it was caused by degrading water conditions through the dataset. As seen in Figure 3, we observed a sharp decrease in all metrics after around the 120th image in both MLP and native runs, which indicates an increase in backscatter. This is consistent with our hypothesis, as the median scaling s_{high} for native is much larger than 1. The images at the end of the dataset are visibly much more turbid than in the beginning; however, we cannot find a classification method based on one image alone, and exploring this topic is outside the scope of this paper.

This means that we can no longer use the first 100 images of D2 as a training subset as their water condition dramatically differs from the rest of the dataset. Therefore, six more MLP models, $Dn_{scatter_x}$ as introduced in Section 4.2, were trained with the same methodology as Dn_{sub_x} to represent a more generalized model, but less so than the mixed Dx. These were run on all datasets, including the unseen Dn_{inv_x}, where $Dn_{inv_x} \bigcup Dn_{scatter_x} = Dn$, and $Dn_{inv_x} \bigcap Dn_{scatter_x} = \emptyset$, for $n \in \{1, 2, 4\}$ and $x \in \mathbb{N}$ [22]. We found the method to dramatically improve the performance of D2 for both DA3 and Depth Pro over $D2_{sub100}$ as seen in Figure 3, where the peaks appear in similar locations, but their magnitudes decrease dramatically. On the unseen Dn_{inv_x}, both models achieved similar performance level, with an average improvement of 56%, 59%, 54%, 811%, and 43%, 45%, 37%, 239%, for *RMSE*, *MAE*, *AbsRel*, $\delta 1$, respectively across DA3 and Depth Pro. This indicates strong generalization improvement within close water conditions.

5 Discussion

Although the MLP scale predictor gave significant uplifts in similar locations and water conditions and could have uses with stationary cameras, its results are less accurate than needed for estimating fish size and by extension the health of the fish and coral population. As water conditions can change quickly, a method lowering the cost of obtaining full-frame depth information underwater is still necessary, since data covering all conditions would be required for an all-encompassing model. Our method would also be improved by a dataset with highly accurate depth maps where the camera location is stationary. Furthermore, as the predictors require the

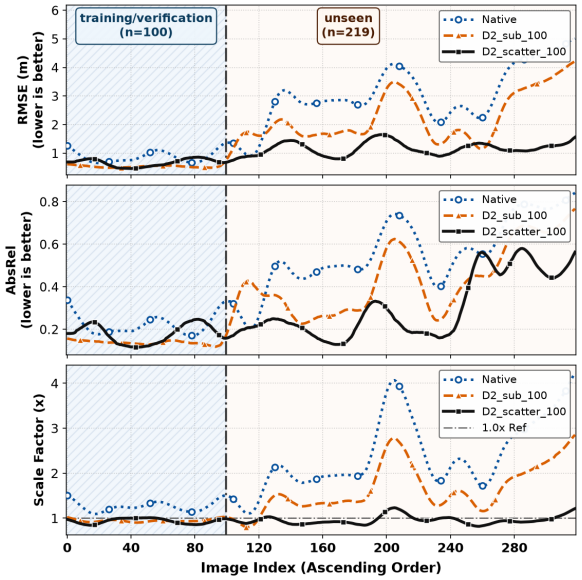


Figure 3: DA3 metrics on D2. The RMSE and AbsRel represent raw values, and the Scale Factor represents the s required to be applied to reach the median scaling results as an upper bound.[3]

water conditions to be nearly identical to the dataset they are trained on, an approach to quantify and classify these conditions based on a single underwater image would also be beneficial. The results we obtained for single-point ground truth scaling are promising with high accuracy. After verification tests in more conditions, this approach is useful in fixed underwater cameras with at least one known depth a priori, autonomous vehicles with a calibrated laser, and mapping the environment in images where a distance is known.

Acknowledgments

Funding and Resources. This work was funded in part by the Research Experiences for Undergraduates Site Engineers for Exploration at UC San Diego (NSF Award 2244123) and the Matthew Henson Fellowship Program supported by the UC HBCU Program. We also acknowledge the CSE department at UCSD and the UCSD IEEE Supercomputing Club at San Diego Supercomputer Center for hosting compute resources for this research.

Declaration of Generative AI and AI-Assisted Technologies. The authors used Gemini (Google, 3.1, accessed August 2026) to create run and conversion scripts and check for potential errors. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Derya Akkaynak and Tali Treibitz. 2019. Sea-Thru: A Method for Removing Water From Underwater Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[2] Shlomi Amitai, Itzik Klein, and Tali Treibitz. 2023. Self-Supervised Monocular Depth Underwater. In *2023 IEEE International Conference on Robotics and Automation (ICRA).* 1098–1104. doi:10.1109/ICRA48891.2023.10161161

[3] Dana Berman, Deborah Levy, Shai Avidan, and Tali Treibitz. 2021. *SQUID- Stereo Quantitative Underwater Image Dataset.* doi:10.5281/zenodo.5744037

[4] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. 2025. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second. In *International Conference on Learning Representations.* https://arxiv.org/abs/2410.02073

[5] John Burns, Donna Delparte, RD Gates, and Misaki Takabayashi. 2015. Integrating structure-from-motion photogrammetry with geospatial software as a novel technique for quantifying 3D ecological characteristics of coral reefs. *PeerJ* 3 (07 2015). doi:10.7717/peerj.1077

[6] Zijie Cai and Christopher Metzler. 2025. Underwater Monocular Metric Depth Estimation: Real-World Benchmarks and Synthetic Fine-Tuning with Vision Foundation Models. arXiv:2507.02148 [cs.CV]

[7] Christopher Crutchfield, Kyle Hu, Nathan Hui, Omair Qazi, Jennifer Loch, Allison Candelmo, Brice Semmens, Curt Schurgers, and Ryan Kastner. 2026. FishSense Lite: A Camera-Based, Single Laser Hardware Framework with Software System for In Situ Fish Length Measurement. doi:10.5281/zenodo.22882398

[8] Luca Ebner, Gideon Billings, and Stefan Williams. 2024. Metrically Scaled Monocular Depth Estimation through Sparse Priors for Underwater Robots. In *2024 IEEE International Conference on Robotics and Automation (ICRA).* 3751–3757. doi:10.1109/ICRA57147.2024.10611007

[9] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/91c56ce4a249fae5419b90cba831e303-Paper.pdf

[10] Renata Ferrari, Mitch Bryson, Tom Bridge, Julie Hustache, Stefan B. Williams, Maria Byrne, and Will Figueira. 2016. Quantifying the response of structural complexity and community composition to environmental change in marine communities. *Global Change Biology* 22, 5 (2016), 1965–1975. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.13197 doi:10.1111/gcb.13197

[11] Mark A. Hixon, Darren W. Johnson, and Susan M. Sogard. 2014. BOFFFFs: on the importance of conserving old-growth age structure in fishery populations. *ICES Journal of Marine Science* 71, 8 (10 2014), 2171–2185. doi:10.1093/icesjms/fst200

[12] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. 2025. Depth Anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647* (2025).

[13] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 11976–11986.

[14] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG] https://arxiv.org/abs/1711.05101

[15] Yelena Randall and Tali Treibitz. 2023. FLSea: Underwater Visual-Inertial and Stereo-Vision Forward-Looking Datasets. arXiv:2302.12772 [cs.RO]

[16] Anna Szafarczyk and Cezary Toś. 2023. The Use of Green Laser in LiDAR Bathymetry: State of the Art and Recent Advancements. *Sensors* 23, 1 (2023). doi:10.3390/s23010292

[17] Chen Wang, Haiyong Xu, Gangyi Jiang, Mei Yu, Ting Luo, and Yeyao Chen. 2024. Underwater Monocular Depth Estimation Based on Physical-Guided Transformer. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–16. doi:10.1109/TGRS.2024.3373904

[18] Emily Wong, Isabella Humphrey, Scott Switzer, Christopher Crutchfield, Nathan Hui, Curt Schurgers, and Ryan Kastner. 2022. Underwater Depth Calibration Using a Commercial Depth Camera. In *Proceedings of the 16th International Conference on Underwater Networks & Systems* (Boston, MA, USA) *(WUWNet '22).* Association for Computing Machinery, New York, NY, USA, Article 22, 5 pages. doi:10.1145/3567600.3568158

[19] Jinghe Yang, Mingming Gong, and Ye Pu. 2025. Physics-Informed Knowledge Transfer for Underwater Monocular Depth Estimation. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 449–465.

[20] Fan Zhang, Shaodi You, Yu Li, and Ying Fu. 2024. Atlantis: Enabling Underwater Depth Estimation with Stable Diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 11852–11861. doi:10.1109/CVPR52733.2024.01126

[21] Hongjie Zhang, Gideon Billings, and Stefan B. Williams. 2026. SPADE: Sparsity Adaptive Depth Estimator for Zero-Shot, Real-Time, Monocular Depth Estimation in Underwater Environments. *IEEE Journal of Oceanic Engineering* (2026), 1–11. doi:10.1109/JOE.2026.3707696

[22] Shuzheng Zhang, Christopher Crutchfield, Sijan Shrestha, Curt Schurgers, and Ryan Kastner. 2026. *Accuracy Data for Underwater In-Air Monocular Depth Model Scalings and MLP Predictors.* doi:10.5281/zenodo.22883321

[3] Graphed with a Gaussian-weighted moving average (window of 25) to improve readability. Training and unseen subsections depicted only apply to $D2_{sub_{100}}$