2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

UDepth: Fast Monocular Depth Estimation for Visually-guided Underwater Robots

Boxiao Yu, Jiayi Wu and Md Jahidul Islam

***Abstract*— In this paper, we present a fast monocular depth estimation method for enabling 3D perception capabilities of low-cost underwater robots. We formulate a novel end-to-end deep visual learning pipeline named UDepth, which incorporates domain knowledge of image formation characteristics of natural underwater scenes. First, we adapt a new input space from raw RGB image space by exploiting underwater light attenuation prior, and then devise a least-squared formulation for coarse pixel-wise depth prediction. Subsequently, we extend this into a domain projection loss that guides the end-to-end learning of UDepth on over 9K RGB-D training samples. UDepth is designed with a computationally light MobileNetV2 backbone and a Transformer-based optimizer for ensuring fast inference rates on embedded systems. By domain-aware design choices and through comprehensive experimental analyses, we demonstrate that it is possible to achieve state-of-the-art depth estimation performance while ensuring a small computational footprint. Specifically, with 70%-80% less network parameters than existing benchmarks, UDepth achieves comparable and often better depth estimation performance. While the full model offers over 66 FPS (13 FPS) inference rates on a single GPU (CPU core), our domain projection for coarse depth prediction runs at 51.5 FPS rates on single-board Jetson TX2s. The inference pipelines are available at *https://github.com/uf-robopi/UDepth*.**

I. Introduction

Underwater depth estimation is the foundation for numerous marine robotics tasks such as autonomous mapping and 3D reconstruction [1], visual filtering [2], tracking and servoing [3], [4], navigation [5], [6], photometry and imaging technologies [7], and more. Unlike terrestrial robots, visually guided underwater robots have very few low-cost solutions for dense 3D visual sensing because of the high cost and domain-specific operational complexities involved in deploying underwater LiDARs [8], [9], RGB-D cameras [10], or laser scanners [11]. Fast monocular depth estimation thus plays a critical role in enabling real-time 3D robot perception and state estimation for important applications such as subsea monitoring and inspection [12], autonomous exploration [13], and companion robotics [14].

Traditional approaches for underwater depth estimation either use active photon counting on Time-of-flight (ToF) cameras [15] or geometric model adaptation on structure light [16], [17]. Standard depth cameras [10] with water-housing and echo-sounders [3] are also common to acquire underwater scene depths up-to-scale. However, accurate sensory integration and backscatter filtering are major challenges in their field deployments. Another class of approaches holistically evaluates the physical properties of the optics for depth estimation; *i.e.*, they approximate the medium transmission rate, background light, and attenuation coefficient following an underwater image formation model (IFM) [18], [19], [20], and subsequently estimate depth for each pixel. These methods work reasonably well on clear/non-turbid water-bodies (*e.g.*, Ocean Type I-II) [21], and are particularly suitable for offline image processing tasks. However, their generalized end-to-end implementations are rather difficult due to the optical parameters' inherent dependency on wavelength and water-body properties [7].

In recent years, deep visual learning-based methods are used to address these problems with remarkable success [22], [23]. These data-driven methods heavily rely on large-scale datasets, hence researchers often synthetically generate training images [24], [25], [26] due to the lack of RGB-D ground truths for underwater scenes. However, such training pipelines cannot fully capture the complex underwater scene geometry and image statistics, restricting their performance and applicability for practical applications. Additionally,

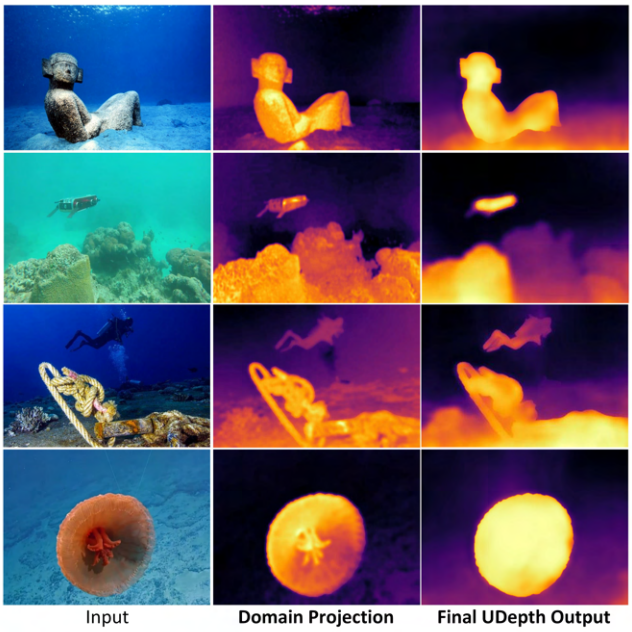

Fig. 1: The proposed UDepth learning pipeline uses the wavelength-dependent attenuation constraints of underwater domain for an intelligent input-space adaptation from raw RGB images. A computationally light domain projection step then approximates a coarse depth map (middle column). The intermediate projections and other pixel-level loss functions guide the end-to-end training of UDepth for fine-grained depth estimations (rightmost column).

The authors are with the Robot Perception and Intelligence (RoboPI) laboratory at the Electrical and Computer Engineering (ECE) department of the University of Florida, USA. Email: boxiao.yu@ufl.edu, wuj2@ufl.edu, jahid@ece.ufl.edu.

their learning objectives are generally adopted from standard terrestrial depth estimation pipelines or coupled with in-air RGB-D image pairs [23], [27]. Consequently, underwater domain knowledge is not incorporated into their learning processes, leading to poor generalization performance.

In this paper, we formulate a robust and efficient end-to-end model named **UDepth**, for fast monocular depth estimation by incorporating underwater domain knowledge into its supervised learning pipeline. We adapt a new input space R**MI** for underwater domain-aware learning, which consists of red channel (R), maximum of green and blue channels (**M**), and grayscale intensities (I). Due to the wavelength-dependent attenuation constraints of underwater light propagation, the relative differences between R and **M** channels embed useful scene depth information while the I channel preserves structural contents of the image. Based on R**MI**, we design a computationally light projection step for pixel-wise **coarse depth prediction** by a least-squared formulation. We further devise a **domain projection loss** to enforce the pixel-wise underwater attenuation constraints in the holistic learning process as well.

The proposed UDepth model architecture consists of the highly efficient MobilenetV2 backbone, a lightweight Vision Transformer (mViT)-based global attention module, and a convolutional regressor for fine-grained depth estimation. The end-to-end inference graph of UDepth has only 15.6M parameters, which is about 80% less than Adabins [28] and 66% less than DenseDepth [29] - two of its closest competitor baselines. As a result, UDepth offers significantly faster inference rates: over 66.7 FPS on a NVIDIA™ RTX 3080 and 13.3 FPS on an Intel™ Core i9-3.50GHz CPU core. With comprehensive domain-aware learning on 9223 RGB-D pairs of natural underwater schenes from USOD10K datset [30], UDepth achieves SOTA performance on standard benchmarks despite having such a light architecture. Moreover, we demonstrate that UDepth offers better generalization performance on arbitrary test cases from Sea-Thru dataset [31] and real-world field experimental data.

Furthermore, we demonstrate that our domain projection module with subsequent filtering can be used for coarse depth prediction on low-power embedded devices. Specifically, it runs at 51.5 FPS on NVIDIA™ Jetson TX2s and 7.92 FPS on Raspberry Pi-4s. More importantly, visual results on field experimental data suggest that its coarse predictions are reasonably accurate and can be used for on-board 3D perception by low-cost underwater robots.

II. Background & Related Work

A. Monocular Depth Estimation Literature

Traditional methods for monocular depth estimation [32], [33] focus on graphical models with hand-crafted geometric priors based on the concepts of structure-from-motion, stereo vision, and multi-view feature matching. These classical methods require multi-view correspondences and significant computational power, yet only generate sparse depth information. Such difficulties have paved the way for powerful deep visual learning-based methods [34], which can learn to infer dense depth maps from single RGB images in an end-to-end manner. The state-of-the-art (SOTA) methods of the modern era for monocular depth estimation can be classified as supervised, semi-supervised, and unsupervised.

1) Supervised Methods: With large-scale paired datasets generated by RGBD cameras, supervised training pipelines of fully-convolutional networks (FCNs) have dominated the SOTA performance over the past decade [34]. The prototypical model architectures adopt progressively upsampled feature representations [35], [36], dilated convolutions [37], or parallel multi-scale feature aggregation [38], [39] to learn fine-grained depth predictions from RGB images. More recent architectures integrate high-resolution representation with multiple lower-resolution *refined* feature maps [40], [41] to improve the local structural details of the prediction. In recent years, various Vision Transformer (ViT) [42]-based spatial *attention blocks* [28], [43] are incorporated to ensure global context awareness. Notable objective functions used by these networks are the reverse Huber loss [44], classification and ordinal regression loss [35], [45], pairwise ranking loss [46], and other adaptive weighting or density losses [28], [47].

2) Semi-supervised and Unsupervised Methods: The unsupervised learning methods avoid the need for large-scale paired data by inferring scene depth from two or more RGB images by exploiting their inherent geometric constraints [48]. Contemporary methods take stereo image pairs [49], [33] or consecutive frames from video [50], [32] as inputs, combined with geometric view reconstruction [51], [52] or camera pose estimation [53] to achieve dense monocular depth prediction. In recent years, attention mechanisms are coupled into such networks to preserve the inherent spatial details of the predicted depth map [54], [55]. Practical loss functions such as left-right disparity consistency loss [33], [56], photometric loss [57], and symmetry loss [58] are generally utilized in these networks as learning objectives. On the other hand, to get more scale information while avoiding costly ground truth, researchers proposed semi-supervised learning methods by combining multiple training stages [59], [60] or training loss terms [61] of supervised and unsupervised learning. The left-right consistency loss [62], mutual distillation loss [63], and teacher-student learning strategies [64] are also introduced into semi-supervised learning pipelines with inspiring results for robust monocular depth estimation.

B. Depth Estimation on Underwater Imagery

SOTA monocular depth estimation methods are not directly applicable off-the-shelf to underwater imagery due to their domain-specific image formation characteristics and scene geometry [7], [65]. Researchers have addressed this in many ways, which can be categorized into active or passive approaches. Prominent *active* approaches use time-correlated single-photon counting on ToF cameras [15] or 3D model adaptation on structure light [16], [17]. On the other hand, *passive* estimation methods incorporate domain knowledge either following a physics-based IFM or from large-scale

data; these models are more popular for their operational simplicity. While atmospheric models were used by early methods [66], contemporary approaches generally adopt the Jaffe-McGlamery underwater IFM [67] to estimate the depth map from the medium transmission map, *i.e.*, the fraction of scene radiance that reaches the camera after absorption and scattering.

The medium transmission map is generally estimated by using the concepts of dark channel prior (DCP) [68], often adapted for underwater scenes as UDCP (underwater DCP) [69], red-channel compensation [70], etc. Blurriness and illumination information are also used to approximate the transmission map and background light first, and then depth maps are estimated by following the IFM [18], [19], [20]. Hence, depth map estimation from the Jaffe-McGlamery model or following the (more comprehensive) revised IFM [7], [31] requires accurate estimation of background light, transmission map, and water-body parameters. However, these parameters are not always known; in fact, they are estimated by using noisy depth priors for image enhancement and color correction [71], [72].

A practical alternative is to apply transfer learning on SOTA deep visual models for fast monocular depth estimation. The idea is to take advantage of large-scale terrestrial RGB-D datasets for pre-training, and then tune the model weights on limited underwater data. The most commonly adopted models are the Monodepth2 [32], AdaBins [28], various U-Net [73] based architectures (*e.g.*, with VGG encoder [74], [75] or ResNet50 encoder [76]), and GANs [22]. Contemporary researchers have also proposed hybrid training pipelines where in-air (*i.e.*, terrestrial) RGB-D pairs drive the supervised training, while the domain (*i.e.*underwater) data is used in a self-supervised manner simultaneously [23], [27]. Despite some early success of these existing approaches, several important aspects of (i) incorporating useful domain knowledge into comprehensive training pipelines, (ii) learning depth prediction from large-scale underwater RGB-D data, and (iii) computational feasibility analysis for robot vision - are not explored in depth in the literature.

III. Proposed Learning Pipeline

A. New Input Space Adaptation: RMI

Although the wavelength-dependency of atmospheric light attenuation and scattering are negligible, they impact underwater image formation significantly. According to the Jaffe-McGlamery underwater IMF [67], an observed image $\mathbf{U}_\lambda(\mathbf{x})$ can be expressed as:

$$\mathbf{U}_\lambda(\mathbf{x}) = \mathbf{I}_\lambda(\mathbf{x}) \cdot t_\lambda(\mathbf{x}) + \mathbf{B}_\lambda \cdot (1 - t_\lambda(\mathbf{x})), \qquad (1)$$

where $\lambda \in \{\text{R,G,B}\}$ is the wavelength component, $\mathbf{I}_\lambda(\mathbf{x})$ is the clear latent image, $\mathbf{B}_\lambda$ denotes the global background light, and $t_\lambda(\mathbf{x})$ represents the medium transmission rate. The $t_\lambda(\mathbf{x})$ is a function of pixel depth $d(x)$ and medium attenuation coefficient β_λ, defined as: $t_\lambda(\mathbf{x}) = e^{-\beta_\lambda \cdot d(\mathbf{x})}$.

Many contemporary works use priors like DCP [68] or UDCP [69] to approximate $t_\lambda(\mathbf{x})$ and estimate coarse scene depth $d(\mathbf{x})$. Since red wavelength suffers more aggressive attenuation underwater, techniques such as underwater light attenuation prior (ULAP) [21] and red-channel compensation [70] can exploit the R channel values to further refine the depth prediction. As illustrated in Fig. 2, the relative differences between $\{\text{R}\}$ and $\{\text{G, B}\}$ channel values encode useful depth information for a given pixel. In this paper, we exploit these inherent relationships and demonstrate that **RMI**$\equiv\{\text{R, }\mathbf{M}{=}max\{\text{G,B}\}\text{, I (intensity)}\}$ is a significantly better input space for visual learning pipelines of underwater monocular depth estimation models.

B. Network Architecture: UDepth Model

As illustrated in Fig. 3, the network architecture of UDepth model consists of three major components: a MobileNetV2-based encoder-decoder backbone, a transformer-based refinement module (mViT), and a convolutional regressor. These components are tied sequentially for the supervised learning of monocular depth estimation.

1) ***MobileNetV2 backbone****:* We use an encoder-decoder backbone based on MobileNetV2 [77] as it is highly efficient and designed for resource-constrained platforms. It is considerably faster than other SOTA alternatives with only a slight compromise in performance, which makes it feasible for robot deployments. It is based on an *inverted residual* structure with residual connections between *bottleneck* layers [77]. The intermediate expansion layers use lightweight depthwise convolutions to filter features as a source of non-linearity. The encoder contains a series of fully convolution layers with 32 filters, followed by a total of 19 residual bottleneck layers. We adapt the last convolutional layer of decoder so that it finally generates 48 filters of 320×480 resolution, given a 3-channel **RMI** input.

2) ***mViT refinement****:* Transformers can perform global statistical analysis on images, solving the problem that traditional convolution models can only handle pixel-level information [78]. Due to the heavy computational cost of Vision Transformers (ViT), we adopt a lighter mViT architecture inspired by [28]. The 48 filters extracted by the backbone are 1×1 convolved and flattened to patch embeddings, which serve as inputs to the mViT encoder. Those are also fed to a 3×3 convolutional layer for spatial refinements. The 1×1 convolutional kernels are subsequently exploited to compute the range-attention maps $\mathbf{R}$, which combines adaptive global information with local pixel-level information from CNN. The other embedding is propagated to a multilayer perceptron head with ReLU activation to obtain a 80-dimensional *bin-width* feature vector $\mathbf{f}_b$.

3) ***Convolutional regression****:* Finally, the convolutional regression module combines the range-attention maps $\mathbf{R}$ and features $\mathbf{f}_b$ to generate the final feature map $\mathbf{f}$. To avoid discretization of depth values, the final prediction of depth map $\mathbf{D}$ is computed by the linear combination of bin-width centers ($\bar{\mathbf{f}_b}$), which is given by: $\hat{d} = \sum_k \bar{\mathbf{f}_{bk}}\, \sigma(\mathbf{R}_k)$.

C. Objective Function Formulation

Pixel-wise depth losses. We use two pixel-wise supervised loss functions: $i)$ the $\mathcal{L}_2$ loss, and $ii)$ a scaled version of

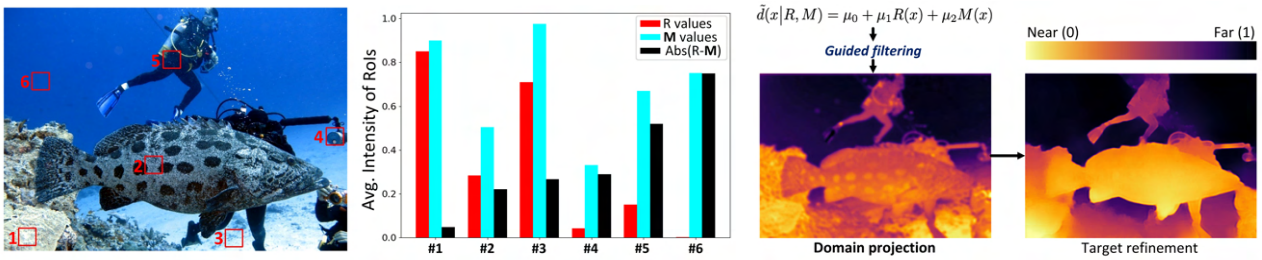

Fig. 2: Red light has the largest wavelength and thus gets attenuated the most underwater [21], [69]. Differences in R and **M**=max{G,B} values vary proportionately with pixel distances, as demonstrated by six 20×20 RoIs selected on the left image. We exploit this domain-specific attenuation constraint with a linear estimator for coarse depth predictions, which can be further filtered for smoothing; an example is shown on the right. Such abstract projection is the basis of our domain projection loss that guides UDepth learning.

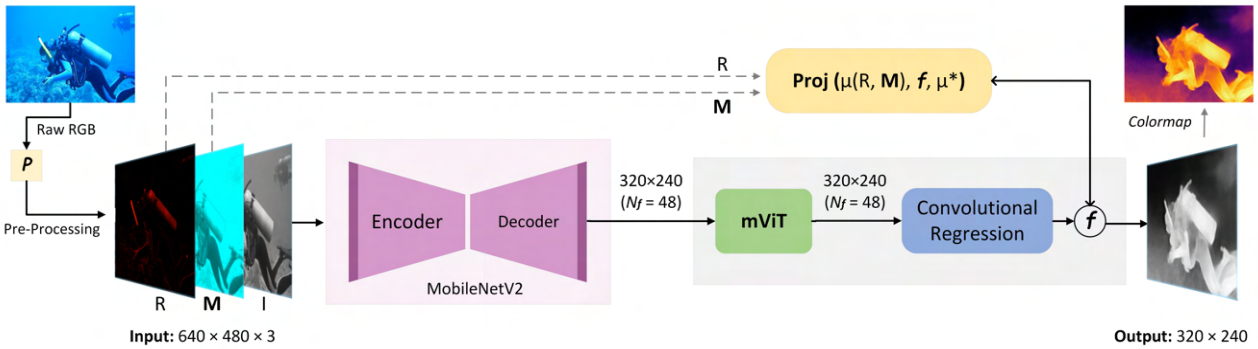

Fig. 3: The end-to-end learning pipeline of our proposed UDepth model is shown. Raw RGB images are first pre-processed to map into **RMI** input space, then forwarded to the MobileNetV2 backbone for feature extraction. Those features are refined by a transformer-based optimizer (mViT), followed by a convolutional regressor to generate the single channel depth prediction. The learning objective involving pixel-wise losses and a domain projection loss is formulated in Eq. 7.

the Scale-Invariant Log (SILog) loss introduced by Eigen *et al.* [40]. These are defined as follows:

$$\mathcal{L}_2 = \mathbb{E}\Big[\big\|d_i - \hat{d}_i\big\|_2\Big], \tag{2}$$

$$\mathcal{L}_{SILog} = \alpha\sqrt{\frac{1}{T}\sum_i g_i^2 - \frac{\lambda}{T^2}(\sum_i g_i)^2}. \tag{3}$$

Here, $g_i = \log \hat{d}_i - \log d_i$ where $\hat{d}$ is the predicted depth, d is the ground truth depth, and T denotes the number of pixels having valid ground truth values. Inspired by [28], we use $\lambda = 0.85$ and $\alpha = 10$ in our implementation.

Domain projection loss. We formulate a novel domain projection loss function based on our input space calibration. Following our discussion in Sec. III-A and Fig. 2, we express the R-**M** relationship with the depth of a pixel x with a linear approximator as follows:

$$\tilde{d}(x|R, M) = \mu_0 + \mu_1 R(x) + \mu_2 M(x). \tag{4}$$

Then, we find the least-squared solution μ^* on the entire RGB-D training pairs, which is over 2.8 billion pixels (9229 images of 640×480 resolution) by optimizing:

$$\mu^* = \arg\min_{\mu} \sum_{i,x} \big\|d_i(x) - \tilde{d}(x|R_i, M_i)\big\|_2^2. \tag{5}$$

We find $\mu^* = [0.464, 0.496, -0.389]$ in our experiments on USOD10K dataset [30] (see Sec. IV-A). Here, our goal is to use the optimal μ-space for regularization, we penalize any pixel-wise depth predictions that violate the underwater image attenuation constraint defined by Eq. 4. We achieve this by the following projection error function:

$$\mathcal{L}_\perp = \mathbb{E}\Big[\Pi_{\mu^*}\big(\tilde{d}_i(R, M) - \hat{d}_i\big)\Big] \tag{6}$$

End-to-end objective. Finally, the end-to-end learning objective of our proposed UDepth pipeline is formulated as:

$$\mathcal{L}_{UDepth} = \lambda_2\, \mathcal{L}_2 + \lambda_S\, \mathcal{L}_{SILog} + \lambda_\perp\, \mathcal{L}_\perp. \tag{7}$$

In Eq. 7, we find the λ-parameters empirically through hyper-parameters tuning. The optimal values used in UDepth training are: $\lambda_2 = 0.3$, $\lambda_S = 0.6$, and $\lambda_\perp = 0.1$.

IV. Experiments

A. Datasets and Evaluation Metrics

*1) **Training and evaluation data**:* We use the USOD10K dataset [30] in our experiments; it contains RGB images and ground truth depth maps for various underwater scenes captured at a pixel resolution of 640×480. The dataset contains 9229 training samples and 1026 testing samples. We also use benchmark images from the Sea-Thru dataset [31] for performance evaluation. Moreover, we test UDepth model on unseen field data collected during oceanic explorations and human-robot collaborative experiments.

*2) **Evaluation metrics**:* We use four standard metrics [40] to compare our method against other SOTA models. These error metrics are defined as follows:

- Mean absolute relative error (Abs Rel): $\frac{1}{n}\sum_p^n \frac{|d_p - \hat{d}_p|}{d}$,
- Squared relative error (Sq Rel): $\frac{1}{n}\sum_d^n \frac{||d_p - \hat{d}_p||^2}{d}$,

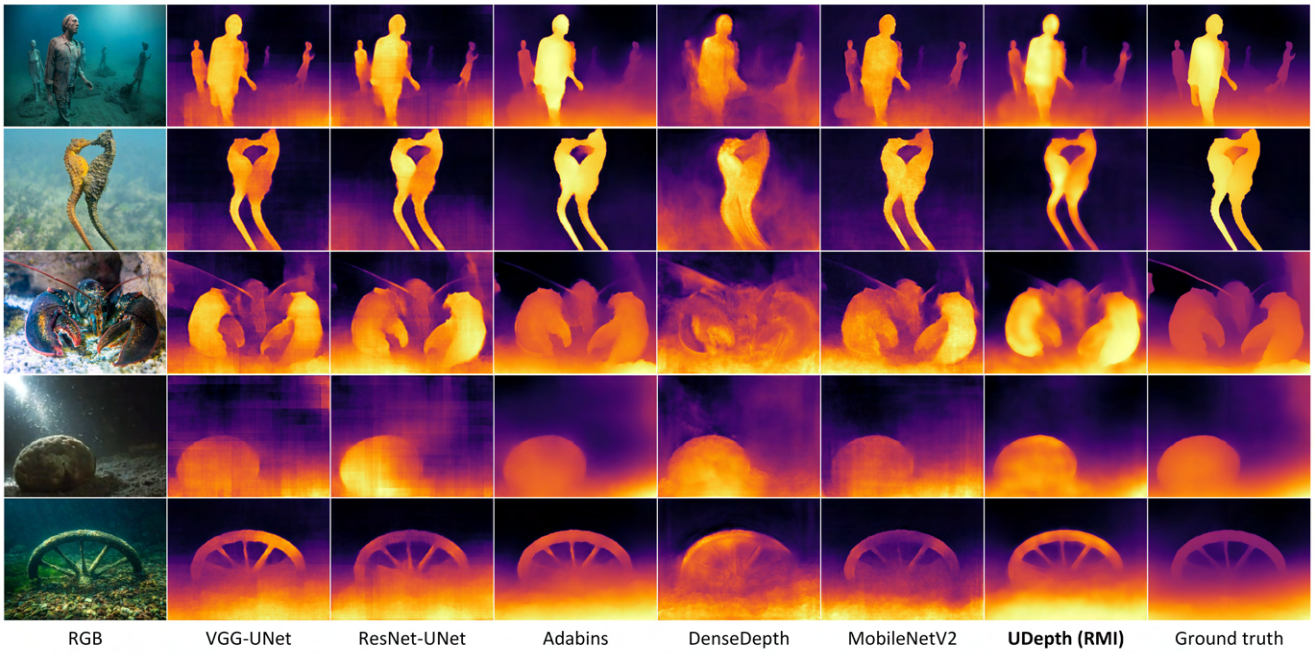

Fig. 4: A few qualitative comparisons are shown for underwater scene depth estimation by UDepth and SOTA monocular depth estimation models on USOD10K test set [30]. As seen, UDepth infers accurate and consistent depth predictions across various waterbody, attenuation levels, and lighting conditions. UDepth offers comparable and often better performance than its closest competitor baseline Adabins, while offering $5\times$ faster inference and memory efficiency.

- Root mean squared error (RMSE): $\sqrt{\frac{1}{n}\sum_{d}^{n}(d_p - \hat{d}_p)^2}$,
- $\log_{10}$ error: $\frac{1}{n}\sum_{d}^{n}|\log_{10}(d_p) - \log_{10}(\hat{d}_p)|$,

Where d_p is a pixel in depth image d, $\hat{d}_p$ is a pixel in the predicted depth image $\hat{d}$, and n is the total number of pixels in d (as well as $\hat{d}$).

B. Implementation Details

UDepth training is supervised by RGB-D image pairs of natural underwater scenes; the RGB color images are pre-processed to prepare **RMI** input images and the respective D channel depth maps serve as ground truth. A total of 7383 images in the USOD10K training dataset are used for training and the remaining 1846 images are used for validation. We use Pytorch libraries [79] to implement its learning pipeline; AdamW [80] is used as the optimizer with an initial learning rate of 0.0001 and exponential decay adjustment with a multiplicative factor of 0.9. With a batch size of 4, Udepth training takes about 5 minutes per epoch on a single node with NVIDIA™ RTX 3080 GPU. It has about 15.5M parameters in total: 3.5M for the MobileNetV2 encoder, 10.9M for the decoder, and 1.1M for the mViT-based refiner and the convolutional regressor combined. For visual image generation, we apply guided filtering [81] to the raw UDepth output for smoothing; here, we used binary saliency map [82] as the guided mask in our implementation.

C. Qualitative and Quantitative Evaluation

For baseline performance comparison, we consider the following five models that are widely used for supervised learning of monocular depth estimation: VGG-UNet [74], [73], ResNet-UNet [76], [73], Adabins [28], DesneDepth [29], and MobileNetV2 [77]. We train these models by following their recommended settings on the same pipeline and *train-validation* data splits as UDepth. The qualitative performances of all these models for some samples are illustrated in Fig. 4, while the quantitative results are listed in Table I-II.

The two major findings of our experimental analyses are as follows: (i) All models exhibit consistently better results across almost all metrics when trained on the **RMI** space instead of raw RGB inputs, which validates our contribution to domain-aware input space adaptation. (ii) The proposed UDepth model outperforms VGG-UNet, ResNet-UNet, DesneDepth, and MobileNetV2 regardless of using **RMI** or RGB input space. On some metrics, the SOTA model Adabins achieves better scores and generate more accurate visual results. However, Adabins is a significantly heavier model with 78M+ parameters compared to UDepth, which has only 15.5M parameters. Hence, our design choices enable UDepth to achieve comparable and often better depth estimation performance than Adabins, despite being over $5\times$ more efficient (at only 20% computational cost).

The qualitative results corroborate our analyses; as the visual comparisons of Fig. 4 illustrate, the output of UDepth and Adabins are more accurate and consistent with the underwater scene geometry. Udepth does a particularly better job at removing background regions and predicting foreground layers up to scale. Moreover, UDepth demonstrates much better generalization performance compared to other SOTA models, as evident from the results in Table II. Here, we use the D3 (reef) scenes from the Sea-Thru dataset [31] for testing only. While Adabins offers good results, UDepth achieves significantly better results across all metrics on unseen test cases. In comparison, DenseDepth model shows slightly better performance on unseen underwater images;

TABLE I: Quantitative comparison for underwater scene depth estimation performance by UDepth and other SOTA models on USOD10K test set [30]. Here, lower scores represent better performance for all metrics in consideration.

Model	Abs Rel	Sq Rel	RMSE	$\log_{10}$
VGG-UNet (RGB)	0.939	0.183	0.150	0.199
VGG-UNet (**RMI**)	0.886	0.180	0.150	0.197
ResNet50-UNet (RGB)	0.875	0.171	0.152	0.200
ResNet50-UNet (**RMI**)	0.847	0.164	0.154	0.204
AdaBins (RGB)	0.617	0.097	0.109	0.154
AdaBins (**RMI**)	0.613	0.090	0.108	0.153
DenseDepth (RGB)	1.178	0.254	0.168	0.219
DenseDepth (**RMI**)	1.128	0.241	0.170	0.220
MobileNetV2 (RGB)	0.858	0.156	0.134	0.183
MobileNetV2 (**RMI**)	0.790	0.144	0.136	0.184
UDepth (RGB)	0.809	0.147	0.129	0.176
UDepth (**RMI**)	0.681	0.123	0.143	0.188

TABLE II: Quantitative performance comparison on Sea-thru (D3) dataset [31] (same models and metrics in consideration as Table I).

Model	Abs Rel	Sq Rel	RMSE	$\log_{10}$
VGG-UNet (RGB)	1.402	0.680	0.439	0.359
VGG-UNet (**RMI**)	1.271	0.563	0.399	0.337
ResNet50-UNet (RGB)	1.272	0.563	0.397	0.335
ResNet50-UNet (**RMI**)	1.206	0.613	0.420	0.352
AdaBins (RGB)	1.314	0.663	0.441	0.356
AdaBins (**RMI**)	1.258	0.617	0.426	0.346
DenseDepth (RGB)	1.073	0.448	0.366	0.312
DenseDepth (**RMI**)	1.092	0.470	0.370	0.312
MobileNetV2 (RGB)	1.196	0.565	0.408	0.337
MobileNetV2 (**RMI**)	1.170	0.547	0.400	0.329
UDepth (RGB)	1.304	0.653	0.440	0.355
UDepth (**RMI**)	1.153	0.514	0.388	0.321

however, with 42.6M parameters, it is about 3× computationally heavier. UDepth's superior performance, particularly for background segmentation and depth continuity on unseen natural underwater scenes are illustrated in Fig. 5.

D. Coarse Depth Estimation by Domain Projection

We further compare the computational efficiency of UDepth with Adabins and DenseDepth in Table III. UDepth is 3-5 times memory efficient and offers 4.4-6.8 times faster inference rates, with over 66 FPS inference on a single RTX-3080 GPU and over 13.5 FPS on CPUs. More importantly, we can extend our domain projection step with guided filtering [81] for fast *coarse depth prediction* on low-power embedded devices. We performed thorough evaluations on field experimental data, which suggest that these abstract predictions are reasonable approximations of natural underwater scene depths. As shown in Fig. 6, the filtered domain projections embed useful 3D information about the scenes to facilitate high-level decision-making by visually-guided underwater robots. The end-to-end domain projection and filtering module generates depth maps at 51.5 FPS on NVIDIA™ Jetson TX2s and 7.92 FPS on Raspberry Pi-4s.

TABLE III: Comparison of model parameters, model sizes, and inference rates for three best performing models: on a CPU (Intel™ Core i9-3.50GHz) and a GPU (NVIDIA™ RTX 3080).

Model	# Params	Memory	FPS (CPU)	FPS (GPU)
Adabins	78.0 M	313.8 MB	1.98	33.70
DenseDepth	42.6 M	178.3 MB	3.06	52.34
UDepth	15.6 M	62.7 MB	13.50	66.23

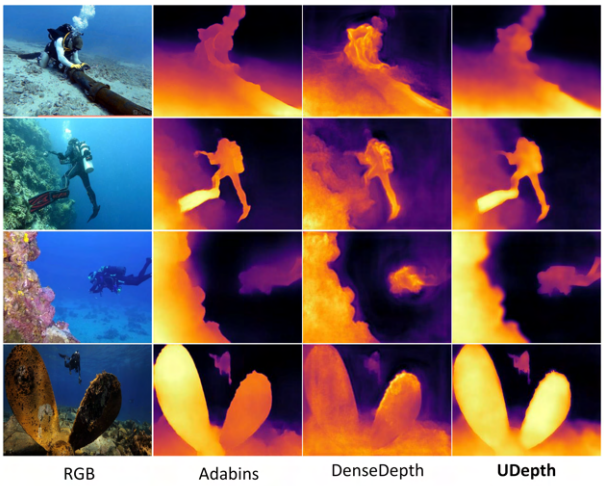

Fig. 5: A few qualitative comparisons for generalization performance of UDepth with its two best competitors: Adabins and DenseDepth are shown. These underwater test images are randomly selected from various benchmark datasets [83], [84].

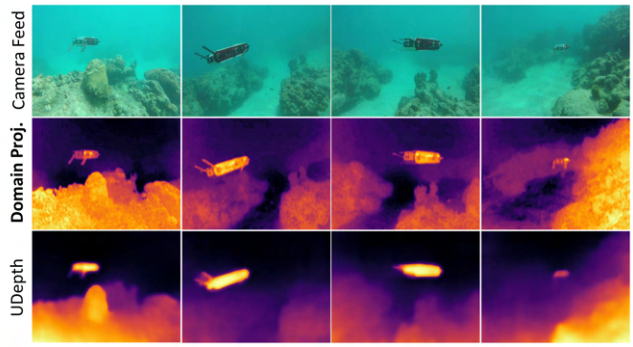

Fig. 6: A few demonstrations of fast coarse depth prediction by our domain projection step are shown in comparison with the final UDepth estimations. The domain projection facilitates an abstract yet reasonably accurate depth prior, which can be used for fast decision-making by underwater robots.

V. Conclusion

Fast monocular depth estimation can facilitate real-time 3D perception capabilities of autonomous underwater robots. In this work, we propose a deep supervised learning pipeline that includes: (i) a domain-aware input space adaptation based on underwater light attenuation characteristics of light propagation; (ii) a least-squared formulation of the attenuation constraints for domain projection of coarse underwater scene depth; and (iii) an efficient deep visual model named UDepth, which can be trained by that domain projection loss and other pixel-level losses for fine-grained monocular depth estimation. The UDepth model is designed with MobileNetV2 backbone and a Transformer-based optimizer to be able to learn an efficient and robust solution for embedded devices. Experimental results show that with only 20% computational cost, UDepth offers comparable and often better depth estimation performance than SOTA models on benchmark datasets and arbitrary test cases. The domain projection also provides a fast solution for low-powered robot deployments. In the future, we plan to extend UDepth's capabilities toward self-supervised depth estimation and image enhancement tasks for real-time underwater robot vision.

REFERENCES

[1] S. T. Digumarti, G. Chaurasia, A. Taneja, R. Siegwart, A. Thomas, and P. Beardsley, "Underwater 3d capture using a low-cost commercial depth camera," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, IEEE, 2016.

[2] M. J. Islam, Y. Xia, and J. Sattar, "Fast Underwater Image Enhancement for Improved Visual Perception," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 2, pp. 3227–3234, 2020.

[3] M. Roznere and A. Q. Li, "Underwater monocular image depth estimation using single-beam echosounder," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1785–1790, 2020.

[4] B.-Y. Raanan, J. Bellingham, Y. Zhang, M. Kemp, B. Kieft, H. Singh, and Y. Girdhar, "Detection of unanticipated faults for autonomous underwater vehicles using online topic models," *Journal of Field Robotics*, vol. 35, no. 5, pp. 705–716, 2018.

[5] M. Xanthidis, N. Karapetyan, H. Damron, S. Rahman, J. Johnson, A. O'Connell, J. M. O'Kane, and I. Rekleitis, "Navigation in the presence of obstacles for an agile autonomous underwater vehicle," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 892–899, IEEE, 2020.

[6] S. Rahman, A. Q. Li, and I. 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)*, pp. 1861–1868, IEEE, 2019.

[7] D. Akkaynak and T. Treibitz, "A Revised Underwater Image Formation Model," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6723–6732, 2018.

[8] G. Zhou, C. Li, D. Zhang, D. Liu, X. Zhou, and J. Zhan, "Overview of underwater transmission characteristics of oceanic lidar," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8144–8159, 2021.

[9] D. McLeod, J. Jacobson, M. Hardy, and C. Embry, "Autonomous inspection using an underwater 3d lidar," in *2013 OCEANS-San Diego*, pp. 1–8, IEEE, 2013.

[10] H. Lu, Y. Zhang, Y. Li, Q. Zhou, R. Tadoh, T. Uemura, H. Kim, and S. Serikawa, "Depth map reconstruction for underwater kinect camera using inpainting and local image mode filtering," *IEEE Access*, vol. 5, pp. 7115–7122, 2017.

[11] A. Palomer, P. Ridao, J. Forest, and D. Ribas, "Underwater laser scanner: Ray-based model and calibration," *IEEE/ASME Transactions on Mechatronics*, vol. 24, no. 5, pp. 1986–1997, 2019.

[12] F. Shkurti, A. Xu, M. Meghjani, J. C. G. Higuera, Y. Girdhar, P. Giguere, B. B. Dey, J. Li, A. Kalmbach, C. Prahacs, *et al.*, "Multi-domain monitoring of marine environments using a heterogeneous robot team," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1747–1753, IEEE, 2012.

[13] Y. Girdhar, P. Giguere, and G. Dudek, "Autonomous adaptive exploration using realtime online spatiotemporal topic modeling," *The International Journal of Robotics Research*, vol. 33, no. 4, pp. 645–657, 2014.

[14] M. J. Islam, J. Mo, and J. Sattar, "Robot-to-robot relative pose estimation using humans as markers," *Autonomous Robots*, vol. 45, no. 4, pp. 579–593, 2021.

[15] A. Maccarone, A. McCarthy, X. Ren, R. E. Warburton, A. M. Wallace, J. Moffat, Y. Petillot, and G. S. Buller, "Underwater depth imaging using time-correlated single-photon counting," *Optics express*, vol. 23, no. 26, pp. 33911–33926, 2015.

[16] M. Massot-Campos and G. Oliver-Codina, "Optical sensors and methods for underwater 3d reconstruction," *Sensors*, vol. 15, no. 12, pp. 31525–31557, 2015.

[17] X. Liu, Y. H. Tan, and B. M. Chen, "Underwater depth map estimation from video sequence with graph cuts," in *2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*, pp. 1–6, 2018.

[18] Y.-T. Peng, X. Zhao, and P. C. Cosman, "Single underwater image enhancement using depth estimation based on blurriness," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 4952–4956, 2015.

[19] J. Raihan A, P. E. Abas, and L. C. De Silva, "Depth estimation for underwater images from single view image," *IET Image Processing*, vol. 14, no. 16, pp. 4188–4197, 2020.

[20] H.-H. Chang, C.-Y. Cheng, and C.-C. Sung, "Single underwater image restoration based on depth estimation and transmission compensation," *IEEE Journal of Oceanic Engineering*, vol. 44, no. 4, pp. 1130–1149, 2018.

[21] W. Song, Y. Wang, D. Huang, and D. Tjondronegoro, "A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration," in *Pacific Rim Conference on Multimedia*, pp. 678–688, Springer, 2018.

[22] P. Hambarde, S. Murala, and A. Dhall, "Uw-gan: Single-image depth estimation and image enhancement for underwater images," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[23] 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, 2019.

[24] Q. Zhao, Z. Zheng, H. Zeng, Z. Yu, H. Zheng, and B. Zheng, "The synthesis of unpaired underwater images for monocular underwater depth prediction," *Frontiers in Marine Science*, p. 1305, 2021.

[25] Q. Zhao, Z. Xin, Z. Yu, and B. Zheng, "Unpaired underwater image synthesis with a disentangled representation for underwater depth map prediction," *Sensors*, vol. 21, no. 9, p. 3268, 2021.

[26] J. Cui, L. Jin, H. Kuang, Q. Xu, and S. Schwertfeger, "Underwater depth estimation for spherical images," *Journal of Robotics*, vol. 2021, 2021.

[27] H. Gupta and K. Mitra, "Unsupervised single image underwater depth estimation," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 624–628, IEEE, 2019.

[28] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4009–4018, 2021.

[29] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv:1812.11941*, 2018.

[30] L. Hong, X. Wang, G. Zhang, and M. Zhao, "USOD10K: A New Benchmark Dataset for Underwater Salient Object Detection." Online: github.com/LinHong-HIT/USOD10K. Accessed: 09-09-2022.

[31] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1682–1691, 2019.

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

[33] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 270–279, 2017.

[34] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, "Monocular depth estimation based on deep learning: An overview," *Science China Technological Sciences*, vol. 63, no. 9, pp. 1612–1627, 2020.

[35] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2002–2011, 2018.

[36] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[37] S. Qiao, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3997–4008, 2021.

[38] J.-H. Lee, M. Heo, K.-R. Kim, and C.-S. Kim, "Single-image depth estimation based on fourier domain analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 330–339, 2018.

[39] J. Li, R. Klein, and A. Yao, "A two-streamed network for estimating fine-scaled depth maps from single rgb images," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3372–3380, 2017.

[40] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[41] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, pp. 2650–2658, 2015.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," *Advances in neural information processing systems*, vol. 30, 2017.

[43] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021.

[44] L. Zwald and S. Lambert-Lacroix, "The berhu penalty and the grouped effect," *arXiv preprint arXiv:1207.6868*, 2012.

[45] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3174–3182, 2017.

[46] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao, "Structure-guided ranking loss for single image depth prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 611–620, 2020.

[47] J.-H. Lee and C.-S. Kim, "Multi-loss rebalancing algorithm for monocular depth estimation," in *European Conference on Computer Vision*, pp. 785–801, Springer, 2020.

[48] C. Zhao, Y. Tang, and Q. Sun, "Unsupervised monocular depth estimation in highly complex environments," *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–10, 2022.

[49] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European conference on computer vision*, pp. 740–756, Springer, 2016.

[50] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1851–1858, 2017.

[51] J. Watson, M. Firman, G. J. Brostow, and D. Turmukhambetov, "Self-supervised monocular depth hints," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2162–2171, 2019.

[52] M. Poggi, F. Tosi, and S. Mattoccia, "Learning monocular depth estimation with unsupervised trinocular assumptions," in *2018 International conference on 3d vision (3DV)*, pp. 324–333, IEEE, 2018.

[53] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2485–2494, 2020.

[54] R. Dai, Y. Gao, Z. Fang, X. Jiang, A. Wang, J. Zhang, and C. Zhong, "Unsupervised learning of depth estimation based on attention model and global pose optimization," *Signal Processing: Image Communication*, vol. 78, pp. 284–292, 2019.

[55] X. Xu, Z. Chen, and F. Yin, "Multi-scale spatial attention-guided monocular depth estimation with semantic enhancement," *IEEE Transactions on Image Processing*, vol. 30, pp. 8811–8822, 2021.

[56] N. Yang, R. Wang, J. Stuckler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 817–833, 2018.

[57] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 340–349, 2018.

[58] W. Zhou, E. Zhou, G. Liu, L. Lin, and A. Lumsdaine, "Unsupervised monocular depth estimation from light field image," *IEEE Transactions on Image Processing*, vol. 29, pp. 1606–1617, 2019.

[59] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin, "Single view stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 155–163, 2018.

[60] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 484–500, 2018.

[61] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6647–6655, 2017.

[62] A. J. Amiri, S. Y. Loo, and H. Zhang, "Semi-supervised monocular depth estimation with left-right consistency using deep neural network," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 602–607, IEEE, 2019.

[63] J. Baek, G. Kim, and S. Kim, "Semi-supervised learning with mutual distillation for monocular depth estimation," *arXiv preprint arXiv:2203.09737*, 2022.

[64] J. Cho, D. Min, Y. Kim, and K. Sohn, "A large rgb-d dataset for semi-supervised monocular depth estimation," *arXiv preprint arXiv:1904.10230*, 2019.

[65] P. L. Drews, E. R. Nascimento, S. S. Botelho, and M. F. M. Campos, "Underwater depth estimation and image restoration based on single images," *IEEE computer graphics and applications*, vol. 36, no. 2, pp. 24–35, 2016.

[66] X. Ding, Y. Wang, J. Zhang, and X. Fu, "Underwater image dehaze using scene depth estimation with adaptive color correction," in *OCEANS 2017 - Aberdeen*, pp. 1–5, 2017.

[67] X. Wu and H. Li, "A simple and comprehensive model for underwater image restoration," in *2013 IEEE International Conference on Information and Automation (ICIA)*, pp. 699–704, 2013.

[68] K. He, J. Sun, and X. Tang, "Single Image Haze Removal using Dark Channel Prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.

[69] 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.

[70] A. Galdran, D. Pardo, A. Picón, and A. Alvarez-Gila, "Automatic red-channel underwater image restoration," *Journal of Visual Communication and Image Representation*, vol. 26, pp. 132–145, 2015.

[71] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, "Color balance and fusion for underwater image enhancement," *IEEE Transactions on image processing*, vol. 27, no. 1, pp. 379–393, 2017.

[72] K. A. Skinner, J. Zhang, E. A. Olson, and M. Johnson-Roberson, "Uwstereonet: Unsupervised learning for depth estimation and color correction of underwater stereo imagery," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7947–7954, IEEE, 2019.

[73] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241, Springer, 2015.

[74] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[75] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*, pp. 239–248, IEEE, 2016.

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[77] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[78] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[79] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[80] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[81] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 6, pp. 1397–1409, 2012.

[82] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[83] M. J. Islam, R. Wang, and J. Sattar, "SVAM: Saliency-guided Visual Attention Modeling by Autonomous Underwater Robots," in *Robotics: Science and Systems (RSS)*, (NY, USA), 2022.

[84] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376–4389, 2019.