2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

StereoAdapter: Adapting Stereo Depth Estimation to Underwater Scenes

Zhengri Wu[2*] **Yiran Wang**[3*] **Yu Wen**[2*] **Zeyu Zhang**[1*†] **Biao Wu**[2] **Hao Tang**[1‡]

[1]School of Computer Science, Peking University [2]AI Geeks [3]The University of Sydney

[*]Equal contribution. [†]Project lead. [‡]Corresponding author: bjdxtanghao@gmail.com

***Abstract*— Underwater stereo depth estimation provides accurate 3D geometry for robotics tasks such as navigation, inspection, and mapping, offering metric depth from low-cost passive cameras while avoiding the scale ambiguity of monocular methods. However, existing approaches face two critical challenges: (i) parameter-efficiently adapting large vision foundation encoders to the underwater domain without extensive labeled data, and (ii) tightly fusing globally coherent but scale-ambiguous monocular priors with locally metric yet photometrically fragile stereo correspondences. To address these challenges, we propose StereoAdapter, a parameter-efficient self-supervised framework that integrates a LoRA-adapted monocular foundation encoder with a recurrent stereo refinement module. We further introduce dynamic LoRA adaptation for efficient rank selection and pre-training on the synthetic UW-StereoDepth-40K dataset to enhance robustness under diverse underwater conditions. Comprehensive evaluations on both simulated and real-world benchmarks show improvements of 6.11% on TartanAir and 5.12% on SQUID compared to state-of-the-art methods, while real-world deployment with the BlueROV2 robot further demonstrates the consistent robustness of our approach.**

I. Introduction

Stereo depth estimation provides accurate *metric* 3D perception from low-cost passive binocular cameras, making it foundational in robotics tasks such as navigation, manipulation, and inspection, without the scale ambiguity of monocular methods [1], [2]. In underwater scenarios, precise depth is equally critical for AUV/ROV-based mapping, inspection, ecological monitoring, and archaeological surveys, where geometric reliability directly impacts autonomy and safety [3], [4].

Recent approaches explore complementary fusion of monocular priors and stereo geometry: TiO-Depth combines monocular and binocular cues within a unified self-supervised framework [5], while Stereo Anywhere injects priors from monocular vision foundation models (VFMs) to enhance generalization in challenging conditions [6]. However, underwater imaging severely violates terrestrial photometric assumptions due to attenuation, scattering, and refraction, introducing a strong domain shift [7]. This presents two key challenges: (i) adapting large vision encoders to the underwater domain in a parameter-efficient manner, and (ii) tightly fusing globally coherent yet scale-ambiguous monocular priors with locally metric but photometrically fragile stereo matches.

We address these issues by designing a self-supervised pipeline that leverages the strengths of both paradigms.

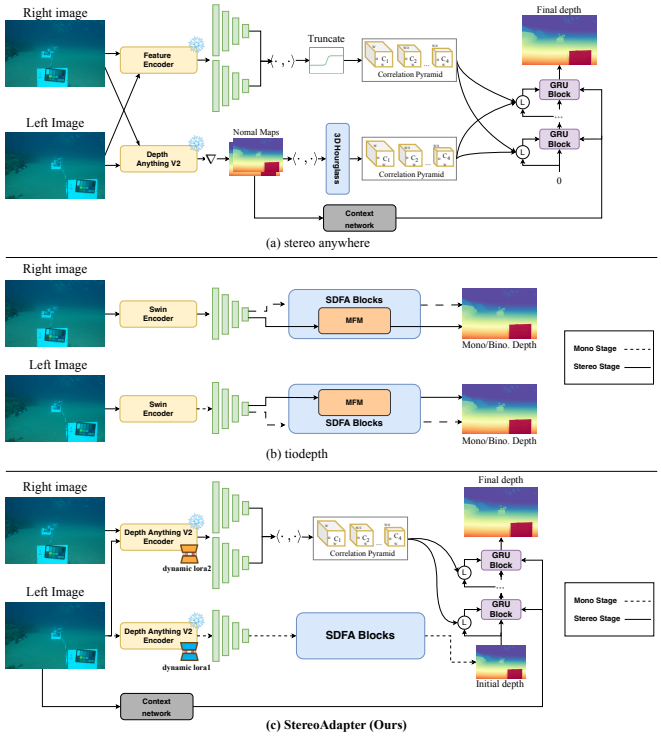

Fig. 1. **Comparison with baselines:** (a) TiO-Depth (b) Stereo Anywhere (c) StereoAdapter (Ours).

Specifically, we (i) *adapt* rather than replace a pre-trained VFM encoder to underwater appearance via lightweight low-rank (LoRA) modules, and (ii) treat the adapted monocular output as a prior to initialize and guide an iterative stereo matcher, which refines local metric depth purely under self-supervision. This yields a robust and deployable framework, even under severe underwater domain shifts.

To this end, we propose **StereoAdapter**, a parameter-efficient self-supervised framework combining a LoRA-adapted monocular foundation encoder with a recurrent stereo refinement module. *Architecturally*, multi-scale features from a Depth-Anything-style encoder (adapted via LoRA) provide a coarse disparity prior, which is fused with a correlation volume in a GRU-style updater to recover full-resolution depth (Fig. 1c). *Training-wise*, we employ dynamic LoRA to automatically select efficient ranks per layer and consolidate them post-training. The system is optimized

with monocular prior guidance, photometric reconstruction, occlusion-aware masking, and smoothness regularizationall without dense labels. On the data side, we synthesize the **UW-StereoDepth-40K** dataset using Unreal Engine 5 to provide high-fidelity underwater stereo, simulating varied attenuation, scattering, particles, and baselines to reflect real-world ROV conditions.

We conduct extensive experiments across synthetic and real-world underwater benchmarks including TartanAir [8] and SQUID [9], achieving RMSE improvements of 6.11

Contributions of this work are:

- We propose **StereoAdapter**, a parameter-efficient self-supervised framework that combines LoRA-adapted monocular priors with stereo refinement.
- We introduce a dynamic LoRA training strategy and the synthetic **UW-StereoDepth-40K** dataset to improve underwater robustness.
- We validate our method on both simulated and real-world benchmarks, demonstrating consistent improvements over prior state-of-the-art approaches.

II. Related Work

a) Deep Stereo Matching and Self-supervised Learning: Deep learning has transformed stereo matching with end-to-end cost-volume networks such as GC-Net [10] and PSMNet [11]. The field later pivoted to iterative refinement with RAFT-Stereo [12], influencing IGEV-Stereo [13] and sparking explorations of Transformer designs [14]. To boost cross-domain generalization, monocular foundation depth modelsMiDaS [15], DPT [16], and Depth Anything/Depth Anything V2 [17], [18]show strong zero-shot transfer; these priors have been fused with stereo geometry, as in Stereo Anywhere [6], and adapted specifically for stereo in FoundationStereo [19] and DEFOM-Stereo [20].

The scarcity of ground-truth depth has spurred self-supervised learning. For monocular estimation, Zhou et al. [21] learn from video by jointly estimating depth and ego-motion, while Monodepth [22] and Monodepth2 [23] exploit stereo leftright photometric consistency. Subsequent refinements introduce depth hints [24], reverse distillation cycles between monocular and stereo learners [25], and reveal reciprocal relations between self-supervised stereo and monocular training [26]. TiO-Depth [5] further unifies monocular and binocular tasks in a single framework, underscoring their complementarity. Collectively, these methods lessen reliance on labels and are particularly valuable in data-scarce settings such as underwater environments.

b) Domain Adaptation for Stereo Matching: Despite advances in stereo matching, cross-domain generalization remains challenging [27], [28]. While methods like masked representation learning [29] show promise for terrestrial scenes, they struggle with extreme domain shifts. Existing datasets primarily focus on terrestrial environments: synthetic datasets (Scene Flow [30], HyperSim [31]) and real benchmarks (KITTI [32], Middlebury [33], Booster [34]) provide comprehensive coverage but cannot capture underwater challenges.

Parameter-efficient fine-tuning offers a promising solution for domain adaptation with limited data. LoRA [35] and its variants [36], [37], [38], [39] demonstrated that large models can be adapted with minimal trainable parameters, particularly effective when target domain data is scarce. Recent vision applications [40], [41] show that adaptive parameter distribution improves adaptation efficiency in challenging domains.

c) Underwater Depth Estimation: Underwater imaging violates fundamental assumptions of standard vision algorithms due to wavelength-dependent attenuation, backscattering, and refraction [7]. Early underwater datasets [9], [42] lacked stereo depth annotations. UW-Stereo [**?**] provided 29,568 synthetic stereo pairs, but the sim-to-real gap persists.

Recent underwater depth estimation methods address these challenges through various strategies. UWStereo [43] proposed comprehensive adaptation with style, semantic, and disparity modules. UWNet and Fast-UWNet [44] developed attention mechanisms for real-time underwater processing. However, these methods require extensive underwater data or complex adaptation pipelines.

The combination of severe domain shift and data scarcity in underwater scenarios motivates our approach: leveraging foundation models' robustness, self-supervised learning's adaptability, and parameter-efficient fine-tuning through LoRA [45]. This enables effective underwater depth estimation by adapting pre-trained models with minimal underwater-specific data, bridging the gap between terrestrial knowledge and underwater applications.

III. The Proposed Method

A. Overview

StereoAdapter is a self-supervised pipeline for underwater stereo depth that leverages monocular estimation to guide stereo prediction as shown in Fig. 2 (a). In Stage-1, we use Depth Anything V2 [18] and adapt its encoder to underwater scenes via LoRA [35], then decode the feature pyramid to obtain a coarse scene disparity. In Stage-2, this disparity initializes the stereo branch and is fused with a cost-volume pyramid inside a recurrent module; progressive iterations yield fine-grained disparities that are converted to depth. Finally, inspired by [45], we incorporate dynamic LoRA to enable continual cross-domain adaptation and tighter coupling of the VFM encoder with stereo disparity prediction as shown in Fig. 2 (b).

B. Architecture

a) Monocular Depth Estimation: We use a pretrained Depth Anything V2 [18] and take the DPT reassemble features $\{F_i\}_{i=1}^{4}$ at $\{H/4, H/8, H/16, H/32\}$. To adapt to underwater imagery with minimal overhead while retaining geometric priors, we insert LoRA [45] into the transformer encoder. The SDFA blocks [46] progressively fuse adjacent scales, and the decoder outputs a discrete disparity volume $V \in \mathbb{R}^{N\times H\times W}$.

For training, the decoder produces a primary volume V_m used for photometric reconstruction following [5]. Using the

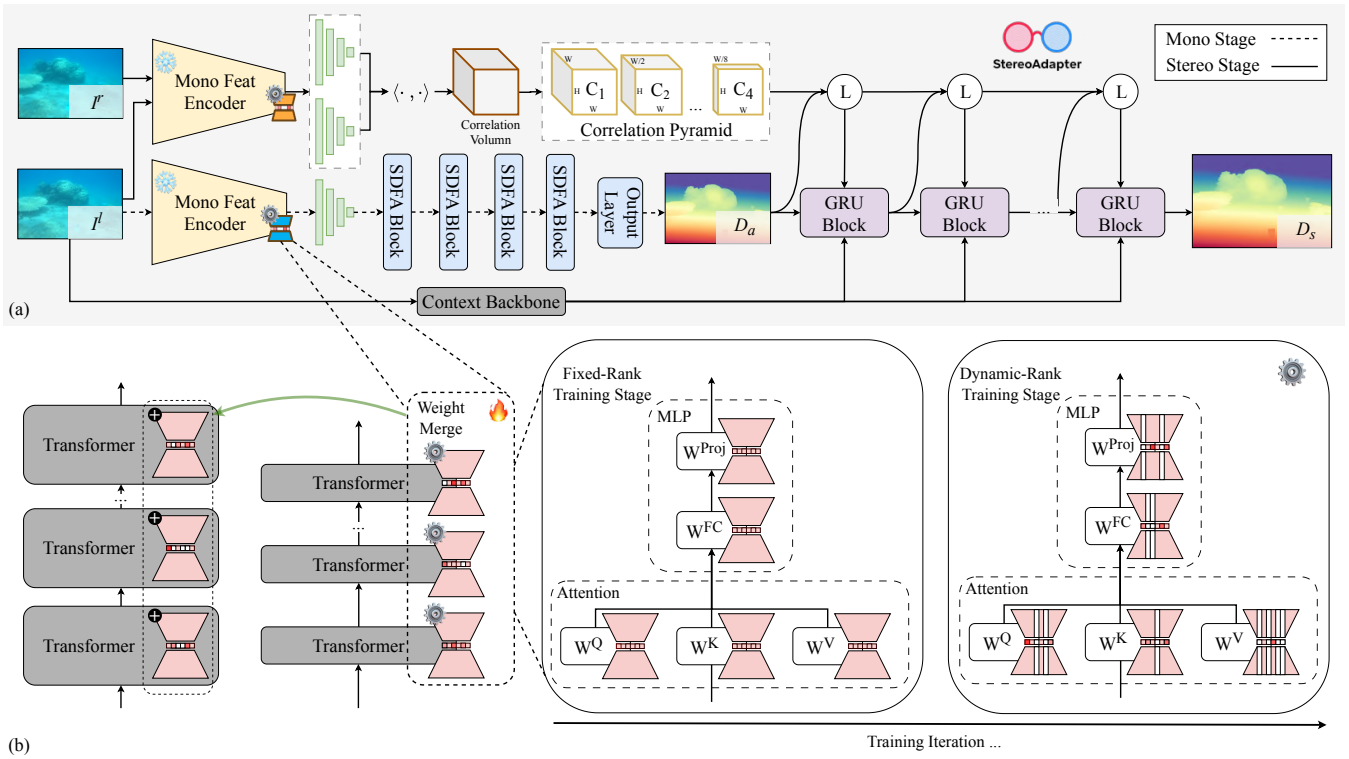

Fig. 2. **Detailed architecture of the StereoAdapter:** (a) two-stage self-supervised training pipeline; (b) update mechanism with LoRA.

discrete depth constraint, we reconstruct $\hat{I}_l$ from V_m^r and I_r. The objective is

$$\mathcal{L}_{mono} = \mathcal{L}_{rec}^{mono} + \lambda_1 \mathcal{L}_{smooth}^{mono}, \tag{1}$$

where $\mathcal{L}_{rec}^{mono}$ compares $\hat{I}_l$ with I_l and $\mathcal{L}_{smooth}^{mono}$ is multi-scale edge-aware smoothness. During this stage, only LoRA and decoder parameters are optimized; the pretrained encoder remains frozen.

b) Correlation Pyramids Building: We construct multi-scale correlation pyramids from stereo features and refine the mono metric predictions using sparse correspondences.

Stereo Correlation Pyramid. Given $\mathbf{f}_L, \mathbf{f}_R \in \mathbb{R}^{C\times H/4\times W/4}$, the 4D correlation is

$$\mathbf{C}(i,j,d) = \langle \mathbf{f}_L(i,j), \mathbf{f}_R(i,j-d)\rangle, \quad d \in [0, D_{\max}]. \tag{2}$$

Average pooling yields a pyramid $\{\mathbf{C}^{(l)}\}_{l=0}^{3}$ for coarse-to-fine refinement.

Hybrid Scale Alignment and Refinement. The mono stage provides metric depths $\mathbf{M}_L^{\text{mono}}, \mathbf{M}_R^{\text{mono}}$. Sparse matches give

$$\mathbf{D}_{\text{sparse}}(p) = \frac{f \cdot b}{d_p}, \quad p \in \mathcal{P}_{\text{matched}}. \tag{3}$$

We first check global scale via

$$\alpha = \frac{1}{|\mathcal{P}_{\text{matched}}|} \sum_{p\in\mathcal{P}_{\text{matched}}} \frac{\mathbf{D}_{\text{sparse}}(p)}{\mathbf{M}_L^{\text{mono}}(p)}. \tag{4}$$

If $|\alpha - 1| < \tau$ (typically $\tau = 0.1$), mono scale is accepted; otherwise we estimate $(\hat{s}, \hat{t})$ by

$$\min_{\hat{s},\hat{t}} \sum_{p\in\mathcal{P}_{\text{matched}}} w_p \left\| \left(\hat{s}\cdot \mathbf{M}_L^{\text{mono}}(p) + \hat{t}\right) - \mathbf{D}_{\text{sparse}}(p)\right\|^2. \tag{5}$$

Aligned depths:

$$\hat{\mathbf{M}}_L^{(0)} = \begin{cases} \mathbf{M}_L^{\text{mono}}, & \text{if } |\alpha - 1| < \tau \\ \hat{s}\cdot \mathbf{M}_L^{\text{mono}} + \hat{t}, & \text{otherwise} \end{cases} \tag{6}$$

Local accuracy is improved by confidence-weighted propagation:

$$\hat{\mathbf{M}}_L(p) = \hat{\mathbf{M}}_L^{(0)}(p) + \sum_{q\in\mathcal{P}_{\text{matched}}} w_{pq}\cdot\left(\mathbf{D}_{\text{sparse}}(q) - \hat{\mathbf{M}}_L^{(0)}(q)\right), \tag{7}$$

with bilateral weights

$$w_{pq} = \exp\left(-\frac{\|p-q\|^2}{2\sigma_d^2}\right) \cdot \exp\left(-\frac{\|\mathbf{I}(p) - \mathbf{I}(q)\|^2}{2\sigma_c^2}\right). \tag{8}$$

c) Stereo Depth Estimation: We refine the mono-scaled initialization using a recurrent stereo module with context features and correlation cues.

Combined Context Encoder. Following DEFOM-Stereo [20], we reuse the mono stages VFM encoder with fused

LoRA for underwater semantics. For $l \in \{4, 8, 16\}$,

$$\mathbf{h}_{\text{ViT}}^{(l)} = \text{Reassemble}_l^{\text{LoRA}}(\mathbf{I}_L), \tag{9}$$

and a lightweight CNN extracts complementary local features:

$$\mathbf{h}_{\text{CNN}}^{(l)} = \text{CNN}_l(\mathbf{I}_L). \tag{10}$$

We align channels and fuse by addition:

$$\mathbf{h}^{(l)} = \text{Conv}_{\text{align}}(\mathbf{h}_{\text{ViT}}^{(l)}) + \mathbf{h}_{\text{CNN}}^{(l)}. \tag{11}$$

The multi-scale $\{\mathbf{h}^{(l)}\}$ initialize and guide the GRU across iterations.

Iterative Disparity Refinement. Starting from $\mathbf{d}^{(1)} = f \cdot b/\hat{\mathbf{M}}_L$, a lookup operator extracts $\mathbf{c}^{(l)}$ from $\{\mathbf{C}^{(l)}\}_{l=0}^{3}$. After a two-layer encoding, we concatenate $\mathbf{c}^{(l)}$, context $\mathbf{g}^{(l)}$, and features of $\mathbf{d}^{(l)}$, then update via ConvGRU:

$$\mathbf{h}^{(l+1)}, \Delta\mathbf{d}^{(l)} = \text{ConvGRU}\Big(\mathbf{h}^{(l)}, \big[\,\mathbf{c}^{(l)}, \mathbf{g}^{(l)}, \phi(\mathbf{d}^{(l)})\,\big]\Big), \tag{12}$$

where $\phi(\cdot)$ extracts disparity features. We upsample as in [12] and convert the final disparity to metric depth.

Joint training strategy. Let $\mathbf{d}^{(L)}$ be the final disparity. The stereo photometric loss is

$$\begin{aligned}\mathcal{L}_{\text{rec}} = \alpha\,\|\tilde{\mathbf{I}}_L - \mathbf{I}'_L\|_1 \\ + (1-\alpha)\,\text{SSIM}\Big(\tilde{\mathbf{I}}_L, \mathbf{I}'_L\Big),\end{aligned} \tag{13}$$

where $\tilde{\mathbf{I}}_L$ is I_R warped by $\mathbf{d}^{(L)}$. Occlusions use mono priors:

$$\begin{aligned}\mathbf{I}'_L = \mathbf{M}_{\text{occ}} \odot \mathbf{I}_L \\ + (1 - \mathbf{M}_{\text{occ}}) \odot \tilde{\mathbf{I}}_L^{\text{mono}},\end{aligned} \tag{14}$$

with $\mathbf{M}_{\text{occ}}$ from $\mathbf{d}^{(1)}$. Monocular guidance regularizes refinement:

$$\begin{aligned}\mathcal{L}_{\text{guide}} =& \|\nabla_x\mathbf{d}^{(1)} - \nabla_x\mathbf{d}^{(L)}\|_1 + \|\nabla_y\mathbf{d}^{(1)} - \nabla_y\mathbf{d}^{(L)}\|_1 \\ &+ \mathbf{M}_{\text{out}} \odot \|\mathbf{d}^{(1)} - \mathbf{d}^{(L)}\|_1,\end{aligned} \tag{15}$$

plus edge-aware smoothness $\mathcal{L}_{\text{smooth}}^{\text{stereo}}$. The final stereo objective is

$$\mathcal{L}_{\text{stereo}} = \mathcal{L}_{\text{rec}}^{\text{stereo}} + \lambda_3\mathcal{L}_{\text{smooth}}^{\text{stereo}} + \lambda_4\mathcal{L}_{\text{guide}}. \tag{16}$$

and we learn a new set of LoRA weights to adapt the VFM encoder from monocular to stereo.

C. Dynamic LoRA

Adaptive Rank Selection Mechanism. We incorporate Low-Rank Adaptation (LoRA) [35] for efficient encoder fine-tuning. For a weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$, traditional LoRA introduces low-rank decomposition matrices $\mathbf{B} \in \mathbb{R}^{d\times r}$ and $\mathbf{A} \in \mathbb{R}^{r\times k}$ to approximate weight updates:

$$h = \mathbf{W}_0\mathbf{x} + \Delta\mathbf{W}\mathbf{x} = \mathbf{W}_0\mathbf{x} + \mathbf{B}\mathbf{A}\mathbf{x}, \tag{17}$$

where $r \ll (\min(d, k))$ represents the adaptation rank.

Inspired by [45], we design a dynamic LoRA that enables adaptive rank optimization by introducing learnable importance weights $\mathbf{w} \in \mathbb{R}^r$ to modulate each rank component:

$$\Delta\mathbf{W}^{t,m} = \sum_{i=1}^{r} \mathbf{w}_i^{t,m}\mathbf{B}_i^{t,m}\mathbf{A}_i^{t,m}, \tag{18}$$

a formulation draws inspiration from the Singular Value Decomposition (SVD) [38], [45], [47], [48], [49], [50], allowing the model to emphasize the most relevant subspace directions for adaptation.

Sparsity-Regularized Optimization. The importance weights $\mathbf{w}$ are jointly optimized with the low-rank matrices using gradient descent. To prune redundant components, we apply an ℓ_1 sparsity regularization:

$$\mathcal{L}_{train}^{t} = \mathcal{L}_{sup}^{t} + \lambda \sum_{m=1}^{M} \|\mathbf{w}^{t,m}\|_1, \tag{19}$$

where $\mathcal{L}_{sup}$ denotes the supervised learning objective, and λ controls the regularization strength.

Since ℓ_1 regularization is non-differentiable, dynamic LoRA employs proximal gradient updates with soft-thresholding to enforce sparsity. The importance weights are updated as:

$$\mathbf{w}_i^{t,m} := \mathbb{1}(\big|\hat{\mathbf{w}}_i^{t,m}\big| > \kappa) \cdot (\hat{\mathbf{w}}_i^{t,m} + \text{sign}(\hat{\mathbf{w}}_i^{t,m}) \cdot \kappa), \tag{20}$$

where $\hat{\mathbf{w}}_i^{t,m}$ denotes the post-gradient update weight and κ is the threshold parameter. The threshold increases from 0 to $\kappa_{\max}$, enabling exploration before pruning. Following the X-TAIL protocol, training adopts two stages: a dense phase (first 50% iterations) without thresholding, followed by a sparse phase that prunes redundant ranks.

Continual Weight Integration. After each adaptation phase, non-zero rank components are merged into the base parameters to eliminate inference overhead:

$$\hat{\mathbf{W}} = \mathbf{W}_0 + \Delta\mathbf{W} = \mathbf{W}_0 + \sum_{i=1}^{r} \mathbf{w}_i\mathbf{B}_i\mathbf{A}_i, \tag{21}$$

where only components with significant weights contribute to the update.

D. Data Synthesis

Inspired by UWstereo [**?**], we construct a large-scale synthetic underwater stereo dataset, **UW-StereoDepth-40K**, using Unreal Engine 5 (UE5), which offers physically accurate rendering and stereo-consistent geometry for depth estimation tasks. Figure 3 illustrates the full pipeline.

a) Rendering Pipeline: Leveraging UE5's ray tracing and global illumination, we build four photorealistic underwater scenescoral reefs, shipwrecks, industrial structures, and seabedspopulated with high-resolution 3D assets (e.g., corals, marine flora, robots, vehicles) from professional libraries and photogrammetry scans. Compared to generative models, UE5 guarantees precise left-right view alignment via calibrated stereo cameras, avoiding stochastic inconsistencies and improving correspondence fidelity.

b) Environmental Variation: To promote generalization, we vary camera baselines {4cm, 10cm, 20cm, 40cm}, reflecting ROV platform diversity. We also simulate realistic underwater phenomena including depth-dependent color attenuation, floating particles with physics-based motion, and water surface caustics.

Fig. 3. **Data synthesis pipeline.** Unreal Engine 5 rendering pipeline for UW-StereoDepth-40K dataset.

TABLE I

EVALUATION ON THE TARTANAIR UNDERWATER SUBSET.

Method	Training Set	REL↓	SQ REL↓	RMSE↓	LOG RMSE↓	A1↑	A2↑	A3↑
Zero-Shot								
LEAStereo [51]	Scene Flow	0.1099	1.3898	4.5610	0.2063	0.8929	0.9512	0.9761
PSMNet [11]	Scene Flow	0.0884	0.8699	3.9721	0.1804	0.9122	0.9627	0.9804
AANet [52]	Scene Flow	0.6096	8.3687	13.0542	0.9903	0.2598	0.3451	0.3888
GwcNet [53]	Scene Flow	0.1013	1.2965	4.1829	0.1855	0.9085	0.9612	0.9801
ACVNet [54]	Scene Flow	0.0970	1.1335	3.9985	0.1803	0.9063	0.9612	0.9813
RAFT-Stereo [12]	Scene Flow	0.0814	0.7342	4.0423	0.1703	0.9030	0.9612	0.9832
HSMNet [55]	Scene Flow	0.9856	12.3768	15.2865	4.5961	0.0000	0.0000	0.0000
TiO-Depth [5]	KITTI2012	0.7194	8.6479	13.4635	1.6967	0.0053	0.0096	0.0550
FoundationStereo [19]	FoundationStereo dataset	0.0542	0.6701	2.9644	**0.1358**	0.9302	0.9701	0.9779
Stereo Anywhere [6]	Scene Flow	0.0592	**0.5098**	3.1572	0.1544	0.9442	0.9787	**0.9889**
CREStereo [56]	ETH3D	2.5746	9.8789	8.4526	5.1297	0.4890	0.5732	0.7001
StereoAdapter (Ours)	UW-StereoDepth-40K	**0.0527**	0.5167	**2.8947**	0.1371	**0.9467**	**0.9801**	0.9873
Fine-Tuning								
IGEV-Stereo [13]	5 Datasets* + TartanAir	0.1009	1.6475	4.7107	0.1909	0.8913	0.9515	0.9764
Selective IGEV [57]	5 Datasets* + TartanAir	0.1225	1.5155	4.8742	0.2123	0.8545	0.9375	0.9713
GMStereo [58]	5 Datasets* + TartanAir	0.1561	2.2275	5.9224	0.2432	0.8362	0.9252	0.9651
StereoAdapter (Ours)	TartanAir	0.0519	0.5041	2.8341	0.1330	0.9489	0.9823	0.9897
StereoAdapter (Ours)	UW-StereoDepth-40K + TartanAir	**0.0512**	**0.4987**	**2.7834**	**0.1312**	**0.9512**	**0.9836**	**0.9904**

c) Dataset Construction: We extract stereo pairs from continuous camera trajectories, automatically filtering low-texture or extreme-depth frames (>50m). Each pair is rendered at 1280×720 with ground-truth depth and segmentation. We ensure stereo consistency via SSIM checks, and experts manually inspect for visual fidelity. The final dataset contains 40,000 high-quality pairs spanning diverse underwater scenes for robust stereo model training.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

a) Datasets: For training, we utilize our synthetic dataset UW-StereoDepth-40K, generated using Unreal Engine 5, which consists of 40,000 stereo image pairs from various underwater scenes. For evaluation, we conduct experiments on two underwater stereo datasets. The first is a subset of TartanAir [8] containing 13,583 underwater stereo image pairs from 22 different sequences. The second is the SQUID dataset [9], which comprises 57 stereo pairs captured from four distinct scenes.

b) Evaluation metrics: We adopt standard depth estimation metrics to comprehensively evaluate our method. Following established protocols, we report Relative Error (REL), Squared Relative Error (SQ REL), Root Mean Square Error (RMSE), and Log Root Mean Square Error (LOG RMSE) to assess accuracy. Additionally, we compute threshold accuracy metrics $\delta < 1.25^i$ (denoted as A1, A2, A3 for $i = 1, 2, 3$ respectively) to measure the percentage of pixels with depth predictions within specified error thresholds.

B. Implementation Details

a) Visual Augmentation: We adopt the pre-trained MobileIE [59] to enhance underwater images by removing water-induced distortions such as color shifts, low contrast, and blurcommon issues caused by light absorption and scattering in water. These degradations severely affect feature extraction and recognition in downstream tasks. By restoring visual quality at the input stage without altering semantics, we provide more reliable inputs. MobileIE offers stable enhancement across training and testing, with a lightweight design and fast inference. Its low computational cost makes it ideal for real-time deployment on resource-constrained underwater ROV platforms.

b) Model Details: The encoder of our model is initialized with pre-trained DepthAnything v2-B [18] weights. We employ a two-stage training strategy: 20 epochs, and followed by 40 epochs for stereo depth estimation. A constant learning rate of 1×10^{-4} is used with the AdamW optimizer and a batch size of 8. Our model is trained on the proposed UW-StereoDepth-40K dataset, with evaluation performed on the TartanAir [8] underwater subset and SQUID [9] datasets. All experiments are conducted on an Intel Xeon Platinum 8469C CPU at 2.60GHz, with a single NVIDIA L40 48GB GPU and 64GB of RAM.

C. Main Results

The "5 Datasets*" refers to a combination of five commonly used stereo matching datasets for training, which are Scene Flow [30], Sintel [60], ETH3D [61], InStereo2K [62], CREStereo [56].

Our experiments show that the proposed StereoAdapter, trained on UW-StereoDepth-40K, consistently outperforms existing stereo matching methods on the TartanAir Underwater and SQUID benchmarks. It achieves state-of-the-art zero-shot performance and further gains when fine-tuned with TartanAir and UW-StereoDepth-40K.

Table I shows the result on the TartanAir Underwater subset, where StereoAdapter achieves the lowest REL (0.0527) and RMSE (2.8947) in the zero-shot setting, with the highest accuracy at A1 (94.67%). Fine-tuning with TartanAir further improves all metrics: RMSE drops to 2.7834, and A1A3 reach 95.12%, 98.36%, and 99.04%.

Table II presents the result on the SQUID dataset. StereoAdapter achieves an RMSE of 1.8843, which is reduced to 1.8621 with fine-tuning, along with top accuracy across all δ thresholds: 94.13% (A1), 97.48% (A2), and 98.52% (A3). These results highlight StereoAdapters robustness and generalization for underwater stereo depth estimation, and benefit from pretraining on UW-StereoDepth-40K.

As shown in Fig. 4, StereoAdapter generates substantially more accurate and visually coherent depth maps than baseline methods (for example, better scale estimation for far range areas).

D. Real World Evaluation

a) Platform: We deploy a BlueROV2 equipped with Zed 2i stereo camera, controlled via STM32 and running inference on Jetson Orin NX (16GB). All data is recorded onboard.

TABLE II

ZERO-SHOT EVALUATION ON SQUID DATASET.

Method	Training Set	REL↓	SQ REL↓	RMSE↓	LOG RMSE↓	A1↑	A2↑	A3↑
LEAStereo [51]	Scene Flow	0.5574	3.9434	5.4659	0.4335	0.6512	0.8042	0.8869
PSMNet [11]	Scene Flow	0.5182	7.1404	4.9186	0.5902	0.7139	0.7999	0.8311
AANet [52]	Scene Flow	7.4801	314.1577	34.7612	1.8994	0.0602	0.1087	0.1570
GwcNet [53]	Scene Flow	0.2294	1.2275	3.0003	0.3799	0.7423	0.8517	0.9005
ACVNet [54]	Scene Flow	1.6030	65.6518	10.3828	0.7293	0.7019	0.7925	0.8321
RAFT-Stereo [12]	Scene Flow	0.0831	0.6946	1.9625	0.1441	0.9235	0.9634	0.9835
HSMNet [55]	Scene Flow	0.9772	7.2766	8.2301	4.0887	0.0000	0.0000	0.0000
CREStereo [56]	ETH3D	2.5746	9.8789	8.4526	5.1297	0.4890	0.5732	0.7001
IGEV-Stereo [13]	5 Datasets* + TartanAir	0.0932	1.4685	2.4741	0.1523	0.9346	0.9712	0.9820
Selective IGEV [57]	5 Datasets* + TartanAir	0.0960	0.9617	1.9268	0.1665	0.9171	0.9555	0.9720
GMStereo [58]	5 Datasets* + TartanAir	3.3442	140.3211	18.7829	1.0219	0.5300	0.6076	0.6578
TiO-Depth [5]	KITTI2012	1.3154	11.6828	7.0930	0.8121	0.1753	0.3346	0.5133
FoundationStereo [19]	FoundationStereo dataset	0.1095	0.7012	2.2510	0.1584	0.8995	0.9433	0.9501
Stereo Anywhere [6]	Scene Flow	0.0952	1.1017	2.4317	0.1586	0.9179	0.9605	0.9763
StereoAdapter (Ours)	UW-StereoDepth-40K	0.0806	0.7082	1.8843	0.1469	0.9413	0.9748	0.9852
StereoAdapter (Ours)	UW-StereoDepth-40K + TartanAir	**0.0795**	**0.6823**	**1.8621**	**0.1398**	**0.9428**	**0.9761**	**0.9867**

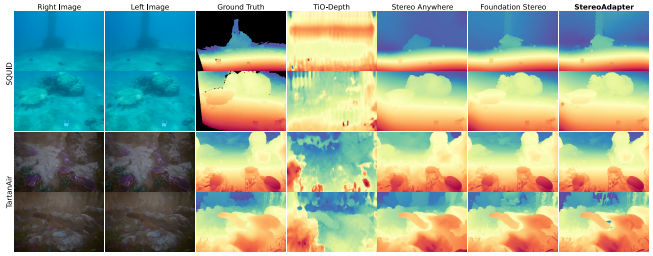

Fig. 4. **Visualization results** of Stereo Depth Estimation Methods on SQUID and TartanAir.

b) Environment: All experiments take place in the same indoor tank with three obstacle setups: dispersed, side-by-side, and clustered. For each, the robot follows three pre-defined trajectories, producing nine stereo sequences. We construct a metrically scaled 3D mesh prior to trials. AprilTag (16h5) detections are used for pose estimation via PnP. For each frame, a reference depth map is rendered in the left-camera frame, masking pixels without valid z-buffer hits.

c) Evaluation Protocol: All methods receive identical rectified stereo pairs. Disparity outputs are converted to metric depth using calibrated focal length f and baseline b. Depth is evaluated in the left-camera frame using only valid pixels. .As in Table III, we report REL, SQ REL, RMSE, LOG RMSE, and A1 ($\delta < 1.25$), averaged over nine sequences. Lower is better (first four), higher is better (A1). We compare against Stereo Anywhere [6], FoundationStereo [19], and TiO-Depth [5]. All methods are evaluated under identical input and pre-processing settings.

E. Ablation Study

We conduct comprehensive ablation studies to assess key design choices in our framework, focusing on: (i) the recurrent refinement module, (ii) the Dynamic LoRA strategy, and (iii) training hyperparameters.

Table IV summarizes the results for the refinement module. Increasing GRU layers and hidden dimensions improves performance, with the 4-layer variant performing best. We adopt a balanced configuration of 3 GRU layers, 128 hidden units, and 32 refinement iterations.

Table V presents the Dynamic LoRA results. We explore adapter rank, κ threshold, and dense epoch ratio. Our final setuprank 16, $\kappa = 0.01$, and 45

TABLE III

REAL WORLD PERFORMANCE

Method	REL↓	SQ REL↓	RMSE↓	LOG RMSE↓	A1↑
Stereo Anywhere [6]	0.0905	1.0467	2.5101	0.1507	0.9120
FoundationStereo [19]	0.1040	0.7363	2.1385	0.1505	0.8961
TiO-Depth [5]	1.5697	13.1487	6.7584	0.8715	0.1525
StereoAdapter (Ours)	**0.0856**	**0.6482**	**1.9690**	**0.1428**	**0.9478**

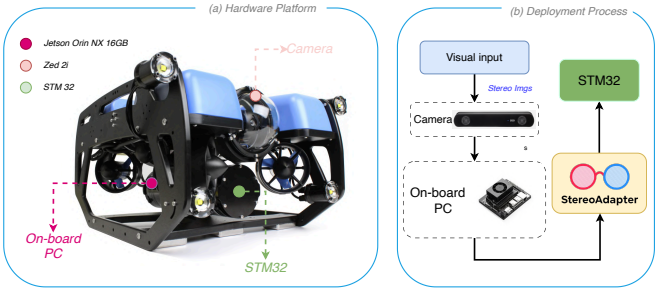

Fig. 5. Hardware and Evaluation Pipeline for Real-World Experiments

Table VI investigates training hyperparameters. Batch sizes of 816 consistently outperform smaller ones, and a learning rate of 1×10^{-4} ensures stable convergence. A multistage schedule (20 epochs for monocular, 40 for stereo) yields the best results (REL = 0.051, RMSE = 2.783) at batch size 8. We adopt this setting as final.

Overall, StereoAdapter achieves state-of-the-art performance across multiple metrics with competitive training and inference efficiency.

V. TEST-TIME EFFICIENCY

TABLE VII

AVERAGE PER-FRAME INFERENCE LATENCY (MS) ON JETSON ORIN NX @ 640×360, BS=1.

Method	**On-board (ms)**
FoundationStereo[19]	1815
Stereo Anywhere [6]	1440
StereoAdapter (Ours)	**1113**

We evaluate on an on-board Jetson Orin NX 16GB (TensorRT), batch size 1, input 640×320. All methods use authors official implementations with identical pre-/post-processing. We report per-frame end-to-end latency (ms). FoundationStereo is slowest, due to its heavy transformer backbone and extensive cost aggregation. Stereo Anywhere reaches 1440 ms with a RAFT-style recurrent module, but time is dominated by two DepthAnything-L monocular passes and a 3D-conv fusion block. In contrast, StereoAdapter is fastest at 1113 ms, using a LoRA-adapted DepthAnything-B encoder solely for feature extraction which is 327 ms faster than Stereo Anywhere and 702 ms faster than FoundationStereo on the same board.

VI. CONCLUSION

We introduce **StereoAdapter**, a parameter-efficient self-supervised framework that couples a LoRA-adapted monoc-

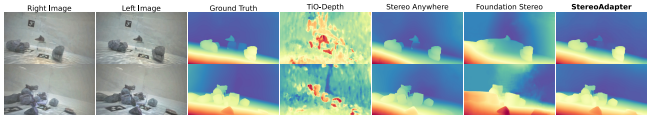

Fig. 6. Generalization of Stereo Methods to Real-World Experiments

TABLE IV
ABLATION ON RECURRENT REFINEMENT MODULE.

GRU layers	Hidden Dimension	Number of Iteration	REL↓	RMSE↓
4	128	32	0.049	2.614
3	256	32	0.048	2.625
3	128	64	0.533	2.8654
2	128	32	0.595	3.024
3	128	32	0.051	2.783

ular foundation encoder with a recurrent stereo refinement module for underwater depth estimation. Dynamic LoRA and pre-training on **UW-StereoDepth-40K** mitigate severe domain shift. On TartanAir and SQUID, StereoAdapter improves accuracy by 6.11% and 5.12% over state of the art, and BlueROV2 deployment confirms robust, low-latency performance. These advances enable practical AUV navigation, infrastructure inspection, and marine ecology monitoring, advancing safer autonomous underwater operations.

TABLE V
ABLATION ON DYNAMIC LORA SETUPS

Rank	κ Threshold	Dense Epoch Ratio	REL↓	RMSE↓
16	0.005	0.5	0.077	3.214
16	0.005	0.45	0.074	3.105
32	0.005	0.5	0.049	2.814
32	0.01	0.5	0.054	2.744
16	0.01	0.45	0.049	2.783

TABLE VI
ABLATION ON TRAINING STRATEGY

Batch Size	Learning Rate	Stage One Epochs	Stage Two Epochs	REL↓/RMSE↓
4	1×10^{-4}	20	20	0.0711/3.183
4	2×10^{-4}	30	60	0.0667/2.951
8	2×10^{-4}	30	30	0.054/2.842
16	1×10^{-4}	20	20	0.052/2.805
8	1×10^{-4}	20	40	0.051/2.783

REFERENCES

[1] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, UK: Cambridge University Press, 2004.

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3354–3361.

[3] S. Zhang, S. Zhao, D. An, J. Liu, H. Wang, Y. Feng, D. Li, and R. Zhao, "Visual slam for underwater vehicles: A survey," *Computer Science Review*, vol. 46, p. 100510, 2022.

[4] F. Nauert and P. Kampmann, "Inspection and maintenance of industrial infrastructure with autonomous underwater robots," *Frontiers in Robotics and AI*, vol. 10, p. 1240276, 2023.

[5] Z. Zhou and Q. Dong, "Two-in-one depth: Bridging the gap between monocular and binocular self-supervised depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9411–9421.

[6] L. Bartolomei, F. Tosi, M. Poggi, and S. Mattoccia, "Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1013–1027.

[7] D. Akkaynak and T. Treibitz, "A revised underwater image formation model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6723–6732.

[8] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "Tartanair: A dataset to push the limits of visual slam," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 4909–4916.

[9] D. Berman, D. Levy, S. Avidan, and T. Treibitz, "Underwater single image color restoration using haze-lines and a new quantitative dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2822–2837, 2020.

[10] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 66–75.

[11] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5410–5418.

[12] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *International Conference on 3D Vision (3DV)*, 2021, pp. 218–227.

[13] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 21 919–21 928.

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

[15] 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*, vol. 44, no. 3, pp. 1623–1637, 2022.

[16] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12 179–12 188.

[17] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[18] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024.

[19] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5249–5260.

[20] H. Jiang, Z. Lou, L. Ding, R. Xu, M. Tan, W. Jiang, and R. Huang, "Defom-stereo: Depth foundation model based stereo matching," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21 857–21 867.

[21] 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*, 2017, pp. 1851–1858.

[22] 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*, 2017, pp. 270–279.

[23] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth prediction," in *The International Conference on Computer Vision (ICCV)*, 2019.

[24] 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 (ICCV)*, 2019, pp. 2162–2171.

[25] F. Aleotti, M. Poggi, and S. Mattoccia, "Reversing the cycle: Self-

supervised deep stereo through enhanced monocular distillation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[26] Z. Chen, X. Ye, W. Yang, Z. Xu, X. Tan, Z. Zou, E. Ding, X. Zhang, and L. Huang, "Revealing the reciprocal relations between self-supervised stereo and monocular depth estimation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15 509–15 518.

[27] F. Zhang, X. Qi, R. Yang, V. Prisacariu, B. Wah, and P. Torr, "Domain-invariant stereo matching networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[28] C. Cai, M. Poggi, S. Mattoccia, and P. Mordohai, "Matching-space stereo networks for cross-domain generalization," in *International Conference on 3D Vision (3DV)*, 2020, pp. 364–373.

[29] Z. Rao, B. Xiong, M. He, Y. Dai, R. He, Z. Shen, and X. Li, "Masked representation learning for domain generalized stereo matching," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[30] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4040–4048.

[31] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind, "Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[32] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[33] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German Conference on Computer Vision (GCPR)*, 2014.

[34] P. Zama Ramirez, F. Tosi, M. Poggi, S. Salti, S. Mattoccia, and L. Di Stefano, "Open challenges in deep stereo: the booster dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[35] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[36] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2023.

[37] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Chao, and M.-H. Chen, "Dora: Weight-decomposed low-rank adaptation," in *Proceedings of the International Conference on Machine Learning*, 2024.

[38] Q. Zhang, M. Chen, A. Bukharin, N. Karampatziakis, P. He, Y. Cheng, W. Chen, and T. Zhao, "Adalora: Adaptive budget allocation for parameter-efficient fine-tuning," in *International Conference on Learning Representations (ICLR)*, 2023.

[39] S. He, L. Ding, D. Dong, M. Zhang, and D. Tao, "Sparseadapter: An easy approach for improving the parameter-efficiency of adapters," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[40] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[41] S. Chang, P. Wang, H. Luo, F. Wang, and M. Z. Shou, "Revisiting vision transformer from the view of path ensemble," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 889–19 899.

[42] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar, "Semantic segmentation of underwater imagery: Dataset and benchmark," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 1769–1776.

[43] X. Ye, J. Zhang, Y. Yuan, R. Xu, W. Zhihui, and H. Li, "Underwater depth estimation via stereo adaptation networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, pp. 4228–4240, 2023.

[44] L. Zhu, Y. Gao, J. Zhang, Y. Li, and X. Li, "Reliable and effective stereo matching for underwater scenes," *Remote Sensing*, vol. 16, no. 23, p. 4570, 2024.

[45] H. Lu, C. Zhao, J. Xue, L. Yao, K. Moore, and D. Gong, "Adaptive rank, reduced forgetting: Knowledge retention in continual learning vision-language models with dynamic rank-selective lora," *arXiv preprint arXiv:2412.01004*, 2024.

[46] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, "Monovit: Self-supervised monocular depth estimation with a vision transformer," in *International Conference on 3D Vision (3DV)*, 2022.

[47] N. Ding, X. Lv, Q. Wang, Y. Chen, B. Zhou, Z. Liu, and M. Sun, "Sparse low-rank adaptation of pre-trained language models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[48] Z. Liu, J. Lyn, W. Zhu, X. Tian, and Y. Graham, "Alora: Allocating low-rank adaptation for fine-tuning large language models," in *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

[49] F. Meng, Z. Wang, and M. Zhang, "Pissa: Principal singular values and singular vectors adaptation of large language models," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2024.

[50] J. Zhang, Y. Zhao, D. Chen, X. Tian, H. Zheng, and W. Zhu, "Milora: Efficient mixture of low-rank adaptation for large language models fine-tuning," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

[51] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, T. Drummond, H. Li, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[52] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1959–1968.

[53] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3273–3282.

[54] G. Xu, J. Cheng, P. Guo, and X. Yang, "Attention concatenation volume for accurate and efficient stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 981–12 990.

[55] H. Zhao, H. Zhou, Y. Zhang, J. Chen, Y. Yang, and Y. Zhao, "High-frequency stereo matching network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1327–1336.

[56] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 263–16 272.

[57] X. Wang, G. Xu, H. Jia, and X. Yang, "Selective-stereo: Adaptive frequency information selection for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 701–19 710.

[58] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, "Unifying flow, stereo and depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[59] H. Yan, A. Li, X. Zhang, Z. Liu, Z. Shi, C. Zhu, and L. Zhang, "Mobileie: An extremely lightweight and effective convnet for real-time image enhancement on mobile devices," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2025.

[60] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 611–625.

[61] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[62] W. Bao, W. Wang, Y. Xu, Y. Guo, S. Hong, and X. Zhang, "Instereo2k: a large real dataset for stereo matching in indoor scenes," *Science China Information Sciences*, vol. 63, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221110870