Wat3R: Underwater 3D Geometry Learning without Annotations

Jiangwei Ren, Xingyu Jiang†, Zijie Song, Wei Xu, Hongkai Lin, Dingkang Liang, and Xiang Bai

Huazhong University of Science and Technology
{jwren,jiangxy998,dkliang,xbai}@hust.edu.cn

Abstract. Estimating 3D geometry in underwater environments presents unique challenges due to light attenuation, scattering, and the absence of large-scale, high-quality 3D annotations. Pioneering methods rely on massive dense annotations that are impractical in underwater settings. In this paper, we propose **Wat3R**, a cross-domain semi-supervised learning framework designed to adapt feed-forward 3D reconstruction models from air to underwater scenes. Uniquely, our method eliminates the need for any annotated underwater data following a teacher-student architecture, that learns robust geometry representations merely on abundant unlabeled real underwater video footage. We also design a cross-view consistency loss that leverages geometric cues from other views to compensate for the information degradation in the current view caused by water attenuation and scattering. Furthermore, considering the lack of comprehensive evaluation benchmarks, we construct **Water3D**, a diverse dataset covering various water bodies and underwater scenarios, designed for geometric task evaluation. Experimental results demonstrate that Wat3R outperforms current state-of-the-art methods in underwater multi-view depth estimation and point cloud reconstruction. The dataset and code are available at https://github.com/LSXI7/Wat3R .

Keywords: Underwater Vision, Geometry Estimation, VGGT, Cross-domain Semi-supervised Learning

1 Introduction

Underwater visual geometry estimation aims to recover the 3D structures, including camera poses, depths, and point clouds, from multi-view underwater imagery. This capability underpins practical applications ranging from underwater robot navigation and obstacle avoidance, to marine mapping, terrain modeling, and underwater archaeology [17,23]. Unlike on-land scenes, underwater environments present unique challenges caused by the light absorption and scattering, as well as view-dependent degradation. These physical difficulties further lead to the critical scarcity of large-scale, high-quality 3D annotations [27] in this domain, posing great challenges in training a well-performing model.

† Corresponding author.

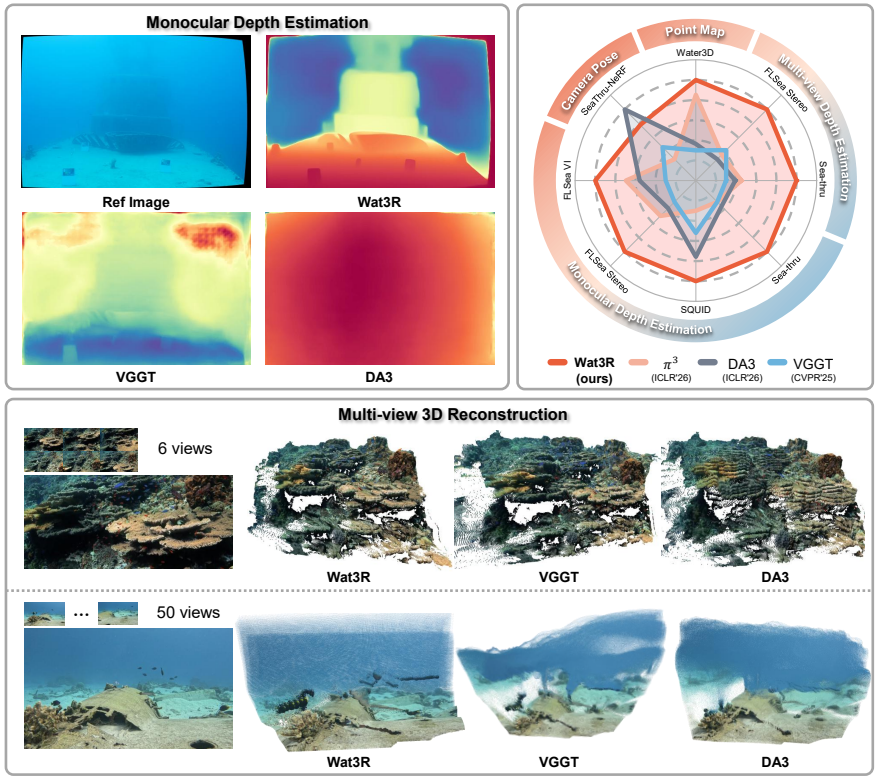
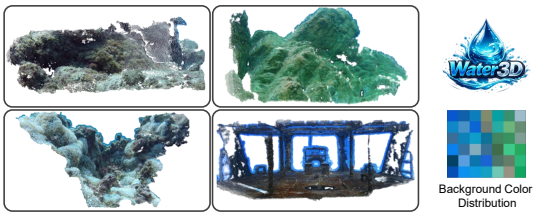


Fig. 1: Wat3R reconstructs from the open-domain underwater images in a feed-forward manner without requiring any underwater 3D annotations. Our Wat3R achieves significant enhancement in both single-view and multi-view tasks. Statistic results also reveal the superior performance of our Wat3R against the SOTA.

In recent years, 3D vision has witnessed a paradigm shift from classical multi-view geometry pipelines to feed-forward neural reconstruction models. Advanced methods like DUSt3R [38], VGGT [37] and their successors [5, 18, 36, 44] have demonstrated remarkable performance in recovering 3D geometry in a single forward pass. These models learn strong geometric priors from massive on-land datasets with dense 3D ground truths, *a.k.a.* camera poses and depths. However, directly deploying these powerful models in underwater scenarios results in poor generalization due to the significant domain shift. And the difficulty in obtaining dense 3D labels for real underwater data also prevents the direct training of effective, generalizable models tailored to underwater environments.

To relieve the above dilemma, we propose **Wat3R**, a semi-supervised framework that adapts VGGT to the underwater domain without any underwater 3D annotations. As shown in Fig. 1, our method enables feed-forward reconstruction across underwater scenes with complex water conditions. Following a teacher-student learning paradigm, we first simulate various underwater degradations



Dataset	#Scenes	Annotations	
		Depth	Pose
SeaThru [3]	5	✓	–
SQUID [4]	4	✓	–
FLSea VI [27]	12	✓	–
FLSea Stereo [27]	5	✓	–
SeaThru-NeRF [19]	4	–	✓
Water3D (Ours)	**42**	✓	✓

Fig. 2: Overview of our constructed Water3D dataset. Left: point cloud visualization and background color distribution of Water3D. Right: comparison of existing underwater datasets in terms of scene scale and available annotations. Our Water3D consists of various underwater conditions with both depth and pose annotations.

on existing annotated on-land datasets (the original training set of VGGT) as the required labeled data, which helps to initialize strong geometric priors for underwater environments. To further refine and generalize the model to a real underwater domain, we also collect a large amount of real underwater video footage as the unlabeled training set. Besides, to address the visual degradation caused by water attenuation and scattering, we introduce the *Cross-view Consistent Loss*, which integrates geometric cues from other views to compensate for the information degradation in the current view. Considering the incomplete annotations and insufficient coverage of existing underwater benchmarks, we construct **Water3D**, which contains a wide spectrum of underwater conditions and scenarios with correct camera poses and depths as shown in Fig. 2. Extensive experiments on public datasets and our Water3D demonstrate that our Wat3R significantly outperforms strong baselines and recent feed-forward alternatives, delivering robust performance even in poor visibility underwater regions.

In summary, our contributions are as follows:

- We propose **Wat3R**, the first semi-supervised VGGT-based framework that generalizes to diverse underwater scenes without requiring underwater 3D annotations. It offers a practical paradigm for geometric learning in other complex environments, just leveraging unlabeled real videos.
- We introduce a *Cross-View Consistent Loss*, which makes it easy to learn geometry cues from heavily degraded regions by aggregating the information from other related views as compensation.
- We construct **Water3D**, an underwater multi-view dataset with comprehensive 3D annotations. It contains diverse underwater conditions.

2 Related Work

2.1 Feed-forward 3D Reconstruction Models

Classical 3D reconstruction is built on multi-view geometry and optimization-based pipelines. Wherein the camera poses and scene structure are recovered through feature matching, epipolar geometry, and global refinement [1, 8–10,

13, 29, 30]. These systems achieve high accuracy and scalability. However, they rely on complex multi-stage pipelines and iterative optimization, leading to high computational cost and limited robustness in challenging imaging conditions like underwater scenes. Recent work reforms 3D reconstruction toward feed-forward inference. DUSt3R [38] initiates this paradigm by directly regressing dense geometry from image pairs without known camera poses. MASt3R [18] strengthens this 3D prior for correspondence and alignment. Subsequent methods [5, 26, 36, 39, 44] progressively extend this framework toward multi-view inference, higher efficiency, and SLAM-oriented applications. VGGT [37] further improves reconstruction accuracy through large-scale training and multi-task joint optimization. More recently, MapAnything [16] expands the input space by incorporating multiple geometric inputs, while Depth Anything 3(DA3) [21] advances large-scale generalization with unified depth–ray modeling. However, the success of these models critically depends on massive labeled 3D datasets. In underwater scenes, obtaining those geometric annotations is difficult due to the poor visibility and unknown underwater imaging, which fundamentally limits the applicability of data-hungry reconstruction models.

2.2 Underwater Vision and Geometry Perception

Underwater visual perception is strongly influenced by the physical image formation process, where light attenuation and scattering significantly degrade visual signals. A large body of work [3, 14, 24] has focused on underwater image restoration and enhancement, aiming to remove color distortion and scattering effects to improve visual quality. Physically grounded underwater imaging models [2] therefore play a central role in characterising underwater visual environments. Several reconstruction and perception frameworks [19, 34, 46] explicitly incorporate these models to jointly reason about scene geometry and underwater light transport. Recent work [42] further shows that incorporating underwater imaging priors can improve LMMs for the understanding of underwater scenes.

Due to the scarcity of large-scale underwater datasets with accurate 3D annotations, several studies explore synthetic data generation using physics-based rendering engines [25, 40] or generative models [22, 47]. However, these approaches are typically designed for stereo or monocular perception tasks and often exhibit domain discrepancies when transferred to real underwater environments. NeRF [19, 31, 49] and 3D Gaussian Splatting methods [20, 35, 43] have also been applied to underwater scene reconstruction. These methods benefit from the compatibility between volumetric rendering formulations and underwater imaging models, enabling joint modeling of scene geometry and light transport. Nevertheless, they generally rely on known camera poses or sparse geometric priors and require optimization-based reconstruction. In contrast, our work focuses on feed-forward multi-view geometry learning in underwater scenes without requiring underwater 3D annotations. For this purpose, we introduce a cross-domain semi-supervised learning framework together with a cross-view consistency loss to address underwater degradation for better training.

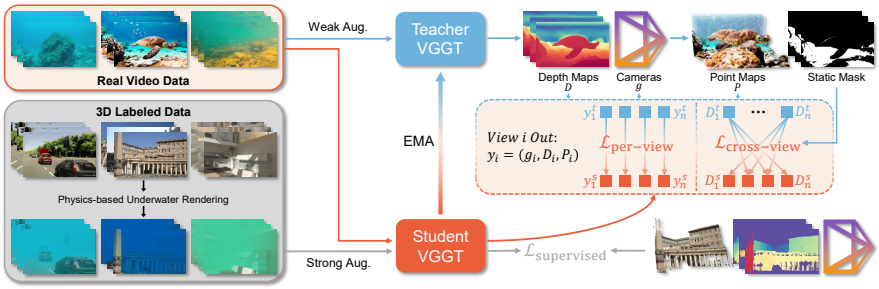


Fig. 3: Overview of our Wat3R framework. Our pipeline follows a Mean Teacher semi-supervised paradigm, where the teacher network produces pseudo-labels for depth, camera parameters, and point maps to supervise the student network. Training leverages labeled synthetic underwater data together with unlabeled real underwater videos, enabling adaptation without underwater 3D annotations. Additional per-view and cross-view consistency losses enforce multi-view geometric coherence.

3 Method

This paper aims to recover camera poses, depths and point clouds from collected underwater views in a single feed-forward pass. To achieve this, several key challenges we may face: *(i) How to train a model under limited or no 3D annotations for underwater scenes. (ii) How to restrain the information loss caused by unknown underwater degradation.* In the following, we will alleviate these concerns by applying cross-domain semi-supervised learning onto the advanced pretrained visual geometry model. Specifically, we first introduce the problem definition and notation of the base model VGGT [37], then describe the overall framework and training strategy, and finally detail our consistency-aware losses.

3.1 Problem Definition and Notation

Given a sequence of underwater RGB frames $(I_i)_{i=1}^N$, the target of our model is to predict a set of geometric attributes y_i for each view i, formulated as:

$$(y_i)_{i=1}^N = f\big((I_i)_{i=1}^N\big)\,, \tag{1}$$

where $y_i = (g_i, D_i, P_i)$ denotes the predicted geometric attributes of view i corresponding to the camera parameters, depth map, and point map, respectively. $g_i \in \mathbb{R}^9$ encodes the camera pose between view i and the first camera, containing the rotation quaternion, translation vector, and field of view. $D_i \in \mathbb{R}^{H\times W}$ denotes the depth map; and $P_i \in \mathbb{R}^{3\times H\times W}$ is the predicted 3D point map expressed in the coordinate frame of the first camera. Following this feed-forward framework, VGGT [37] is known as the pioneering work, which is trained on large-scale 3D-annotated on-land datasets and has shown remarkable performance. Therefore, we chose VGGT as the base model such that it can provide strong geometric priors when adapting to underwater environments. However,

it is impractical to directly train VGGT for underwater adaptation due to the lack of 3D annotated training sets.

3.2 Cross-domain Semi-supervised Learning

Considering the lack of underwater 3D annotations, we investigate a semi-supervised learning framework that adapts a pretrained multi-view geometry model from land to underwater environments. Our key idea is to transfer geometric priors to underwater scenes through consistency-driven teacher–student training. As shown in Fig. 3, the framework follows a Mean Teacher [32] design, where the teacher provides stable pseudo geometric supervision while the student learns from both labeled and unlabeled data.

To be specific, we first simulate various underwater degradations on existing annotated on-land datasets (the original training set of VGGT) as the required labeled data, which helps to initialize strong geometric priors for underwater environments. We then collect a large amount of real underwater video footage as the unlabeled training set, that helps to further refine and generalize the model to the real underwater domain. During training, the student network is updated through backpropagation, while the teacher follows an exponential moving average (EMA) of the student's parameters with the update rule:

$$\theta_t^k = (1-\lambda)\cdot\theta_t^{k-1} + \lambda\cdot\theta_s^k, \tag{2}$$

where θ_s^k and θ_t^k are the student parameters and teacher parameters in iteration k, respectively. λ is the smoothing coefficient.

Synthesized Underwater Data (Labeled). We generate underwater-style training images using a simplified version of the revised underwater imaging model [2]. The image formation is modeled as:

$$I = Je^{-\beta^D z} + B^\infty(1-e^{-\beta^B z}), \tag{3}$$

where J and I denote the clear image and its underwater degradation and z is the depth of the scene. $B^\infty \in \mathbb{R}^3$ denotes the background light, while $\beta^D \in \mathbb{R}^3$ and $\beta^B \in \mathbb{R}^3$ model the attenuation of direct transmission and backscatter, respectively. All three parameters are randomly sampled within $[0, 1]$. The multi-view datasets often contain incomplete or noisy depth annotations, which can create partial water-rendering artifacts when used directly in the image formation model. Therefore, we estimate dense depth maps using the monocular depth model DA3MONO-LARGE [21] for rendering the underwater appearance, while keeping the original geometric annotations as the supervised targets. To control the visible range of each scene, the depth maps are linearly rescaled to a random physical range. The attenuation coefficients are sampled while enforcing the physical constraint that red attenuates faster than green and blue [3]. To simulate spatially varying attenuation, the sampled coefficients are modulated by a smooth random spatial map obtained by repeatedly smoothing Gaussian noise. Several visual results of synthetic underwater can be seen in the bottom left of Fig. 3, which shows the good simulation of different underwater degradations.

Real Underwater Video Data (Unlabeled). To capture the diverse visual conditions of real underwater environments, we construct a large-scale video collection combining a curated scientific dataset with in-the-wild footage. We collect around 10,000 raw underwater videos from public sources, but only 5,504 clips are left after manual filtering. We retain clips with continuous shots, visible static structures, reasonable camera motion, sufficient view overlap, and usable image quality. We remove surface or aerial videos, edited clips with frequent cuts, near duplicates, severe blur or compression, dominant overlays, and clips dominated by open water or moving objects. Frames are sampled every ten frames to reduce temporal redundancy, yielding $359k$ training images in total.

Sequence-level Training Augmentation. As a semi-supervised learning framework, data augmentation is critical for stable training. Following single image based task [6,48], the teacher and student models respectively receive weak and strong augmentations of the same frame. However, the unsupervised training on unlabeled video data often results in limited viewpoint variation. Therefore, sequence-level augmentation is necessary to increase geometric diversity and prevent model collapse. Specifically, we first sample $24-36$ frames and shuffle the order before feeding them to the teacher. For the student branch, the same frames are randomly subsampled to 2–12 images, shuffled again, and each image is rotated by 0°, 90°, 180°, or 270° to avoid pose collapse. We additionally apply color jittering, grayscale conversion, and Gaussian blur.

3.3 Consistency-aware Training Objective

3D reconstruction data generated by SfM pipelines or rendering engines inherently satisfy strong multi-view geometric consistency. To preserve this structural property during training, we introduce additional consistency objectives that encourage coherent geometric predictions. Our training objective consists of three components: a supervised loss following VGGT, a per-view consistency loss using teacher predictions as pseudo-labels, and a cross-view consistency loss enforcing multi-view geometric coherence. The overall training objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{supervised}} + \lambda_p \mathcal{L}_{\text{per-view}} + \lambda_c \mathcal{L}_{\text{cross-view}}. \tag{4}$$

Supervised Loss. The supervised loss is applied when training on labeled synthetic data, which is computed between the student predictions and the ground-truth annotations $\{g^{gt}, D^{gt}, P^{gt}\}$. Here we directly use the loss in VGGT [37]:

$$\mathcal{L}_{\text{supervised}} = \mathcal{L}_{\text{camera}}(g^s, g^{gt}) + \mathcal{L}_{\text{depth}}(D^s, D^{gt}) + \mathcal{L}_{\text{point}}(P^s, P^{gt}), \tag{5}$$

where the depth loss includes regression, confidence, and gradient terms; the point loss consists of regression, confidence, and normal terms; and the camera loss is computed using Huber loss. Details can refer to VGGT [37].

Per-view Geometry Consistency Loss. We use this loss when learning from unlabeled real data. It shares the same formulation as the supervised loss by regarding the predictions of the teacher model, *i.e.,* $\{g^t, D^t, P^t\}$, as pseudo ground

truth. Notably, point maps are projected from depth and camera parameters for more stable geometric supervision [37]. The per-view loss is defined as:

$$\mathcal{L}_{\text{per-view}} = \mathcal{L}_{\text{camera}}(g^s, g^t) + \mathcal{L}_{\text{depth}}(D^s, D^t) + \mathcal{L}_{\text{point}}(P^s, P^t). \tag{6}$$

Cross-view Geometry Consistency Loss. Underwater scattering and attenuation often lead to poor visibility and limited geometric information. To address this challenge, we introduce a cross-view geometry consistency loss that explicitly integrates geometric cues from other views to compensate for the information degradation in the current view. Given N teacher frames, we first obtain a coarse foreground mask M_i^{fg} for each frame i using a simple two-cluster K-means on the depth map, keeping the closer cluster as foreground. For every frame i, we backproject valid teacher pixels into 3D and reproject them into all other teacher views. Define the projected pixel and the visibility indicator as:

$$\begin{aligned} V_{i\to j}(x) &= \mathbf{1}\big(|D_j^t(u_{i\to j}(x)) - D_{i\to j}^t(x)| < \delta \ \wedge \ u_{i\to j}(x) \in \Omega_j\big)\,, \\ u_{i\to j}(x) &= \pi_j\big(\pi_i^{-1}(x, D_i^t(x))\big)\,. \end{aligned} \tag{7}$$

Here, $V_{i\to j}(x)$ indicates whether the pixel x in frame i remains visible and depth-consistent after being projected into frame j. $\mathbf{1}(\cdot)$ returns 1 if it is true. D_i^t denotes the teacher depth for view i. Ω_j is the image domain of view j. $u_{i\to j}(x)$ indicates the projection from view i to view j. Wherein π_j is the projection operator that projects the depth of view j into the world coordinate system using the camera pose, and π_i^{-1} is the back-projection operator. Then the statistic region of view i is computed as the depth-consistent part of its foreground mask:

$$M_i^{\text{static}}(x) = \mathbf{1}\left(\sum_{j\neq i} V_{i\to j}(x) \geq k\right) M_i^{\text{fg}}(x). \tag{8}$$

We use the conservative setting $k = N - 2$ in our experiments, where N denotes the number of teacher frames used to construct the mask. This mask is not an additional supervision source; it only selects reliable pixels for applying cross-view supervision. The foreground term suppresses distant background water regions, while the multi-view consistency term removes dynamic or geometrically unstable regions. When the scene contains mostly open water, severe turbidity, or strong object motion, the mask naturally becomes sparse and therefore prevents the model from enforcing unreliable cross-view constraints.

During student training, we randomly select n frames from the same sequence (in arbitrary order), forming an index set $\mathcal{S}$. For any pair $(i, j) \in \mathcal{S}$, depth from teacher frame i is projected into view j:

$$D_{i\to j}^t(x) = \pi_j\big(\pi_i^{-1}(x, D_i^t(x))\big)\,, \tag{9}$$

and compared with the student depth D_j^s at frame j under the static mask M_i^{static}. The geometry-consistent loss averages over all ordered pairs:

$$\mathcal{L}_{\text{cross-view}} = \frac{1}{|\mathcal{S}|^2} \sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{S}, j\neq i} \mathcal{L}_{\text{L1}}\big(D_j^s,\, D_{i\to j}^t,\, M_j^{\text{static}}\big)\,. \tag{10}$$

Table 1: Multi-view depth estimation. We report results under two evaluation protocols: (1) Shuffle 10-view evaluation, and (2) Full subsequence evaluation (100 frames, ordered). The best and second of each category are masked as **Bold** and Underline, respectively. Shaded rows denote two-stage pipelines, where input images are first enhanced by the corresponding underwater image enhancement method and then fed into VGGT model. Values in (red) indicate the percentage improvement over VGGT.

Methods	Sea-thru [3]				FLSea Stereo [27]			
	Rel↓	δ_1 ↑	$\log_{10}$ ↓	RMSE↓	Rel↓	δ_1 ↑	$\log_{10}$ ↓	RMSE↓
Shuffle 10-view Evaluation (stride=10)								
WaterSplatting [20] 3DV'25	-	-	-	-	0.427	0.415	0.157	1.476
Fast3r [44] CVPR'25	0.277	0.713	0.083	0.631	0.290	0.529	0.141	1.355
MapAnything [16] 3DV'26	0.216	0.800	0.060	0.454	0.146	0.841	0.061	0.798
π3 [39] ICLR'26	0.185	0.909	0.044	0.358	0.139	0.856	0.053	0.837
DA3 [21] ICLR'26	0.187	0.892	0.046	0.333	0.141	0.851	0.056	0.872
VGGT [37] CVPR'25	0.190	0.891	0.047	0.380	0.137	0.849	0.059	0.760
+ Semi-UIR [14] CVPR'23	0.201	0.846	0.052	0.387	0.136	0.861	0.056	0.784
+ PSPL [24] TIP'25	0.209	0.836	0.055	0.419	0.160	0.817	0.063	0.911
Wat3R	**0.167** (+12.1%)	**0.946** (+6.2%)	**0.038** (+19.1%)	**0.290** (+23.7%)	**0.119** (+13.1%)	**0.885** (+4.2%)	**0.048** (+18.6%)	**0.720** (+5.3%)
Full Subsequence Evaluation (100 frames, ordered)								
Fast3r [44] CVPR'25	0.274	0.711	0.081	0.622	0.299	0.521	0.146	1.363
MapAnything [16] 3DV'26	0.227	0.788	0.063	0.488	0.147	0.841	0.062	0.795
π3 [39] ICLR'26	0.187	0.923	0.041	0.358	0.139	0.856	0.053	0.836
DA3 [21] ICLR'26	0.183	0.922	0.040	0.311	0.139	0.855	0.055	0.861
VGGT [37] CVPR'25	0.193	0.900	0.045	0.373	0.136	0.851	0.058	0.759
Wat3R	**0.170** (+11.9%)	**0.953** (+5.9%)	**0.036** (+20.0%)	**0.294** (+21.2%)	**0.120** (+11.8%)	**0.884** (+3.9%)	**0.048** (+17.2%)	**0.715** (+5.8%)

By now, we have introduced the main process of the proposed **Wat3R**, a cross-domain semi-supervised learning framework to generalize VGGT to the challenging underwater domain without requiring any underwater annotations.

4 Experiments

Next, we evaluate our Wat3R on four underwater tasks: multi-view depth estimation (Sec. 4.2), point map estimation (Sec. 4.3), camera pose estimation (Sec. 4.4), and monocular depth estimation (Sec. 4.5).

4.1 Implementation Details

We train the model on 4 NVIDIA RTX 4090 GPUs for 19,200 steps, beginning with a 1,000-step linear warm-up. The peak learning rates are set to 5×10^{-6} for the ViT backbone and 5×10^{-5} for the downstream head. To improve training efficiency and reduce GPU memory consumption, we employ gradient checkpointing and bfloat16 mixed-precision training. The ratio of unlabeled to labeled samples is set to 1:3. For the first 6,400 steps, training is conducted exclusively on labeled data. The unsupervised loss weight λ_u is then gradually increased, reaching its peak value of 0.5 at step 12,800. During training, we randomly sample 2 to 12 images, with the longest side of each image resized to 518 pixels.

Datasets. To ensure applicability to underwater vision scenarios, we conduct experiments on five public datasets together with our own constructed Water3D:

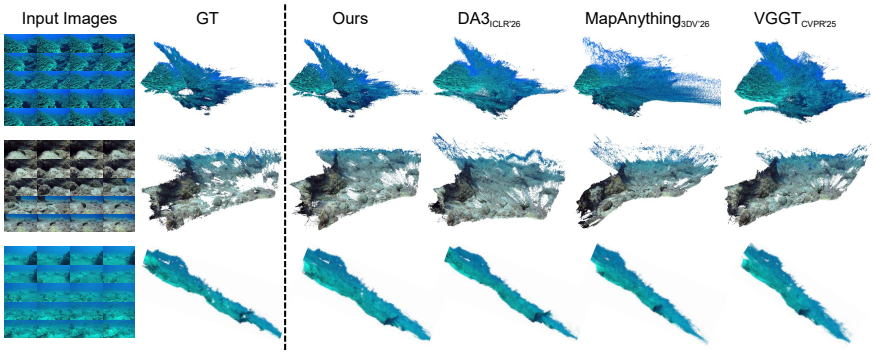


Fig. 4: 3D reconstruction results on our constructed Water3D. DA3 [21], MapAnything [16], VGGT [37] and our Wat3R are used for comparison. Wat3R obtains more complete and physically consistent reconstructions with reduced artifacts.

- **FLSea-VI [27]** contains forward-looking monocular underwater sequences captured in shallow Mediterranean and Red Sea environments. We evaluate on the Coral Table Loop and U-Canyon sequences.
- **FLSea-Stereo [27]** comprises 7,341 synchronized stereo image pairs from 5 shallow-water scenes, including canyon and flat sandy seabed environments, collected from the FLSea-VI dataset in different dives and scenes.
- **SQUID [4]** provides a quantitative underwater stereo dataset consisting of 57 stereo image pairs with stereo-derived ground-truth depth, collected at different seasons, depths, and water types in natural marine environments.
- **Sea-thru [3]** is an underwater RGBD dataset composed of 1,205 images from 5 natural scenes, acquired in two optically different water bodies under natural illumination, spanning depths, water types, and scene structures.
- **SeaThru-NeRF [19]** consists of 5 real underwater scenes annotated with camera poses only. This dataset is originally used for NeRF evaluation.
- **Water3D** consists of 42 real underwater scenes, where 20 are from UVEB [41], and 22 scenes are collected alongside our training data (but not used for training). Camera poses and depths are reconstructed using COLMAP [29], wherein the matcher is changed to the advanced MINIMA [28], which is further fine-tuned with our synthetic underwater data. The final results are manually checked to make sure the correctness of annotations, where only the scenes with stable registration and no obvious collapse or severe outliers are preserved. *See supplement for more details.*

4.2 Multi-view Depth Estimation

We first evaluate the depth estimation from multi-view input. Two video datasets with depth annotations are used, *i.e.,* Sea-thru [3] and FLSea-Stereo [27]. We compare our Wat3R with state-of-the-art models trained on large-scale annotated data. Besides, the underwater Splatting method [20] is also evaluated. To

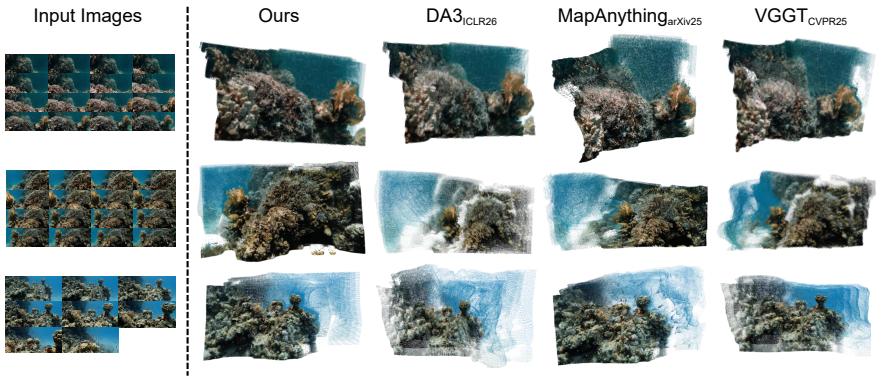


Fig. 5: Qualitative comparison on in-the-wild images with Wat3R, DA3 [21], MapAnything [16], and VGGT [37]. For a fair comparison, all methods take the same input. No post-processing is applied, allowing full point clouds to be visualized.

verify whether the underwater image enhancement (UIE) method is effective for 3D geometry estimation, we also take two enhancement methods into a two-stage pipeline. It first enhances underwater images to make them closer to the on-land domain, then feeds them into a 3D model. Following [37, 38, 44], long sequences are divided into multiple sub-scenes, each containing at most 100 images. And we uniformly sample 10 images from each sequence and randomly shuffle their order for testing. In addition, we follow the video-based depth evaluation protocol of [7] to evaluate full subsequences. We evaluate metric depth by aligning predicted depth maps to the ground-truth scale using a least-squares estimation of scale and shift. Performance is measured using standard depth metrics, and results are reported in Tab. 1. We report absolute relative error (Rel), absolute error in log-scale ($\log_{10}$), root mean squared error (RMSE), and the percentage of inlier pixels δ_i under thresholds of 1.25^i to evaluate overall depth accuracy. The frame order of input video is randomly shuffled during training, so the model does not rely on temporal continuity or video-specific supervision.

Tab. 1 reveals that our model achieves clear and consistent improvements compared to the original VGGT [37]. Specifically, WaterSplatting [20] is a Splatting method designed for underwater scenes. But it relies on initialized point clouds, producing large errors on FLSea-Stereo and even failing on the challenging Sea-thru dataset. π3 [39] and DA3 [21] obtain competitive results due to their improved 3D representations, and their large training set contains some underwater scenes. The shaded rows show that UIE almost brings no improvement for multi-view depth estimation. The enhancement process may easily introduce information loss or view-inconsistent changes, failing to recover new structural cues. In contrast, Wat3R learns underwater-aware geometry directly during training, leading to more reliable 3D perception in challenging underwater scenes. These results highlight our method's ability to achieve superior accuracy and robustness, particularly in challenging underwater multi-view scenarios. These high-fidelity results are visually demonstrated in Fig. 4 and Fig. 5.

Table 2: Performance comparison on the Water3D dataset. The best and second of each category are masked as **Bold** and Underline, respectively. Values in (red) indicate the percentage improvement over VGGT [37].

Methods	Accuracy ↓		Completeness ↓		Overall ↓	
	Mean	Median	Mean	Median	Mean	Median
Fast3r [44] CVPR'25	1.216	0.531	2.444	0.784	1.830	0.658
MapAnything [16] 3DV'26	0.643	0.284	0.666	0.277	0.655	0.281
π3 [39] ICLR'26	0.491	0.168	0.413	0.184	0.452	0.176
DA3 [21] ICLR'26	0.679	0.187	0.528	0.160	0.604	0.174
VGGT [37] CVPR'25	0.486	0.193	0.762	0.191	0.624	0.192
Wat3R (Point)	**0.444** (+8.6%)	**0.148** (+23.3%)	0.409 (+46.3%)	0.165 (+13.6%)	0.427 (+31.6%)	0.157 (+18.2%)
Wat3R (Depth+Cam)	0.446 (+8.2%)	0.162 (+16.1%)	**0.366** (+52.0%)	**0.143** (+25.1%)	**0.406** (+34.9%)	**0.153** (+20.3%)

4.3 Point Map Estimation

We further evaluate the quality of reconstructed multi-view point maps using our underwater dataset Water3D. For each sequence, 20 images are uniformly sampled for evaluation. Following [37, 39, 44], we first perform a coarse Sim(3) alignment using the Umeyama algorithm [33], followed by refinement with the Iterative Closest Point (ICP) algorithm. We report Accuracy, Completeness, and their Overall average as metrics in Tab. 2. The direct point-map output and the reconstruction derived from predicted depth and cameras both achieve strong performance. Their consistent results indicate that adaptation preserves geometric agreement among the point, depth, and camera heads. This is enabled by our consistency-aware objectives, which couple predictions across heads and views even without ground-truth annotations for real underwater videos.

4.4 Camera Pose Estimation

We now evaluate camera pose estimation on the SeaThru-NeRF [19]. Following prior work [37, 38], performance is measured using AUC under different angular thresholds, where AUC is calculated from the minimum of Relative Translation Accuracy (RTA) and Relative Rotation Accuracy (RRA).

As shown in Tab. 3, DA3 [21] achieves the best overall performance across all thresholds. This is mainly due to its depth-ray representation, which provides a minimal yet sufficient formulation for jointly modeling scene geometry and camera motion. Geometry-driven pose inference exhibits natural robustness in underwater environments. Our cross-view consistency learning enables the model to extract complementary geometric cues across views, reducing its sensitivity to water-induced degradation and yielding more stable pose estimates.

4.5 Monocular Depth Estimation

We evaluate monocular depth estimation to assess the model's understanding of underwater degradation. In this setting, no complementary cues from other views are available. Following [22, 47], we evaluate performance on four datasets:

Table 3: Pose estimation AUC at different thresholds on the SeaThru-NeRF [19]. The best and second of each category are masked as **Bold** and Underline, respectively. Values in (red) indicate the percentage improvement over VGGT [37].

Methods	AUC@5°	AUC@15°	AUC@30°
Fast3r [44] CVPR'25	0.040	0.239	0.498
π3 [39] ICLR'26	0.216	0.635	0.809
DA3 [21] ICLR'26	**0.731**	**0.901**	**0.950**
MapAnything [16] 3DV'26	0.074	0.458	0.707
VGGT [37] CVPR'25	0.392	0.707	0.843
Wat3R	0.540(+37.8%)	0.820(+16.0%)	0.906(+7.5%)

Table 4: Monocular depth estimation. The best and second of each category are masked as **Bold** and Underline, respectively. The top three methods (Udepth [46], UW-Depth [12], WaterMono [11]) are underwater-specific models. WaterMono [11] is trained on FLSea VI [27]; its results are shown in gray. Values in (red) indicate the percentage improvement over VGGT [37].

Methods	FLSea VI [27]		FLSea Stereo [27]		SQUID [4]		Sea-thru [3]	
	Rel↓	δ_1 ↑	Rel↓	δ_1 ↑	Rel↓	δ_1 ↑	Rel↓	δ_1 ↑
Udepth [46] ICRA'23	0.212	0.683	0.279	0.550	0.312	0.547	0.166	0.832
UW-Depth [12] ICRA'24	0.330	0.447	0.400	0.427	0.491	0.343	0.184	0.792
WaterMono [11] TIM'25	0.074	0.949	0.206	0.684	0.261	0.544	0.145	0.829
DAv2 [45] NeurIPS'24	0.069	0.958	0.134	0.849	**0.099**	**0.900**	**0.089**	0.940
Fast3r [44] CVPR'25	0.198	0.720	0.213	0.679	0.368	0.522	0.125	0.911
MapAnything [16] 3DV'26	0.086	0.951	0.146	0.837	0.104	0.898	0.104	0.960
π3 [39] ICLR'26	0.081	0.944	0.148	0.831	0.234	0.640	0.113	0.949
DA3 [21] ICLR'26	0.090	0.936	0.154	0.824	0.151	0.802	0.111	0.950
VGGT [37] CVPR'25	0.107	0.888	0.159	0.809	0.194	0.703	0.113	0.942
Wat3R	**0.061** (+43.0%)	**0.971** (+9.3%)	**0.120** (+24.5%)	**0.886** (+9.5%)	0.107 (+44.8%)	0.893 (+27.0%)	0.090 (+20.4%)	**0.976** (+3.6%)

FLSea-VI [27], FLSea-Stereo [27], SQUID [4] and Sea-thru [3], using Absolute Relative Error (Rel) and δ_1 as metrics. The results are summarized in Tab. 4. Although our model is not trained with single-image supervision, it achieves the best performance across four datasets. The cross-view supervision helps separate underwater degradation from scene geometry. The learned geometric prior then transfers to single-view input. Fig. 6 presents qualitative monocular results.

4.6 Ablation Studies

Tab. 5 presents an ablation study to analyze the contribution of each component in our framework. Starting from the pretrained VGGT model as the baseline, we progressively introduce key elements of our method, including real underwater video data, sequence-level augmentation, the proposed cross-view consistency loss $\mathcal{L}_{\text{cross-view}}$, and the static mask M^{static}. For reference, we include a supervised variant trained only on synthetic underwater data to highlight the effect of unlabeled real underwater videos and the proposed consistency constraints.

The results show three clear trends. (1) Training with synthetic underwater rendering already improves performance over the VGGT baseline by partially

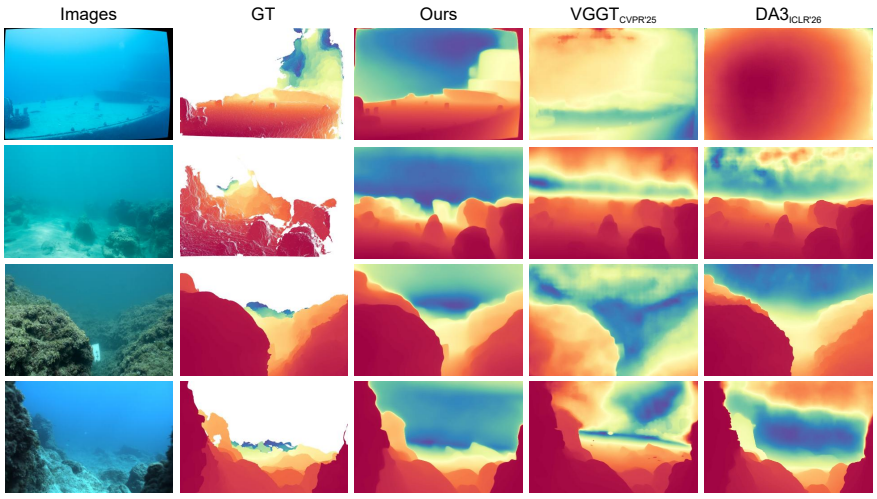


Fig. 6: Monocular depth estimation using Wat3R, DA3 [21], and VGGT [37]. The first two rows are sampled from the SQUID [4] dataset, the third row from FLSea-Stereo [27], and the fourth row from FLSea-VI [27].

Table 5: Ablation studies. The multi-view depth estimation is evaluated. The best and second of each category are masked as **Bold** and Underline, respectively.

	Syn Water	Real Video	Strong Aug.	$\mathcal{L}_{\text{cross-view}}$	M^{static}	Sea-thru		FLSea Stereo	
						Rel↓	δ_1 ↑	Rel↓	δ_1 ↑
VGGT						0.190	0.891	0.137	0.849
	✓					0.173	0.920	0.135	0.860
	✓	✓				0.181	0.906	0.165	0.793
	✓	✓	✓			<u>0.172</u>	0.936	<u>0.126</u>	<u>0.871</u>
	✓	✓	✓	✓		**0.167**	**0.949**	<u>0.126</u>	0.869
	✓		✓	✓	✓	0.174	0.929	0.130	<u>0.871</u>
Wat3R	✓	✓	✓	✓	✓	**0.167**	<u>0.946</u>	**0.119**	**0.885**

bridging the domain gap between terrestrial and underwater imagery. (2) Incorporating real underwater video data together with strong augmentation consistently improves performance, demonstrating that large-scale unlabeled videos provide effective supervision signals for adapting geometry models to underwater scenes. (3) The proposed cross-view consistency loss and static mask further improve performance by enforcing multi-view geometric coherence and suppressing unstable regions caused by scattering or dynamic content. Note that our masking strategy primarily deals with severe underwater degradation and dynamic objects. For simpler and static scenes like Sea-thru, the gains are not significant.

4.7 Robustness Analysis and Failure Cases

To evaluate the robustness of the model under extreme conditions, we report results under progressively stronger underwater degradation on both synthetic and real scenes in Fig. 7. On synthetically degraded DTU [15] dataset, the point-map error of all methods increases as attenuation and scattering intensify, whereas

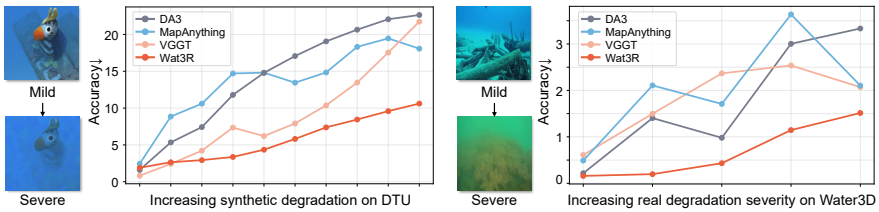


Fig. 7: Robustness and failure case analysis. Left: synthetic underwater degradation with increasing visual corruption. Right: real underwater scenes grouped by difficulty. Wat3R shows more robust to moderate degradation.

Wat3R degrades more gradually and maintains a clear margin over the competing methods. This advantage also extends to real Water3D scenes. Severe visibility loss reduces the stable visual evidence available for geometry recovery. Nevertheless, Wat3R deteriorates more slowly. Its robustness stems from training on diverse synthetic and real underwater conditions, together with consistency objectives that encourage geometry to remain stable under appearance changes.

4.8 Discussion on Possible Limitations

Although Wat3R is designed for underwater geometry reconstruction without requiring underwater annotations, several limitations remain. Handling highly dynamic elements, such as moving marine life or divers, may require additional modeling of temporal dynamics and motion cues. In open-water, highly turbid or very deep scenes, reliable matched structures can become sparse. Our conservative static mask then suppresses unreliable regions, but this also reduces the cross-view learning signal and limits performance in such extreme cases.

5 Conclusion

In this paper, we introduced Wat3R, a first semi-supervised VGGT framework for underwater geometry learning. Wat3R learns robust geometry representations merely on unlabeled real underwater video, without the need for any annotated underwater data. And a cross-view consistency loss is designed to address underwater degradation for better training. Besides, we construct Water3D, the first comprehensive evaluation benchmark containing diverse underwater conditions. Experimental results demonstrate that Wat3R significantly outperforms the state-of-the-art in underwater multi-view geometry estimation.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant U2341227, 62406117, and U25B2078).

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. CACM **54**(10), 105–112 (2011)
2. Akkaynak, D., Treibitz, T.: A revised underwater image formation model. In: CVPR. pp. 6723–6732 (2018)
3. Akkaynak, D., Treibitz, T.: Sea-thru: A method for removing water from underwater images. In: CVPR. pp. 1682–1691 (2019)
4. Berman, D., Levy, D., Avidan, S., Treibitz, T.: Underwater single image color restoration using haze-lines and a new quantitative dataset. IEEE TPAMI (2020)
5. Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V.: Must3r: Multi-view network for stereo 3d reconstruction. In: CVPR. pp. 1050–1060 (2025)
6. Chen, D., Liu, Z., Yang, C., Wang, D., Yan, Y., Xu, Y., Ji, X.: Conformalsam: Unlocking the potential of foundational segmentation models in semi-supervised semantic segmentation with conformal prediction. In: ICCV. pp. 24045–24055 (2025)
7. Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video depth anything: Consistent depth estimation for super-long videos. In: CVPR. pp. 22831–22840 (2025)
8. Crandall, D., Owens, A., Snavely, N., Huttenlocher, D.: Discrete-continuous optimization for large-scale structure from motion. In: CVPR. pp. 3001–3008 (2011)
9. Crandall, D.J., Owens, A., Snavely, N., Huttenlocher, D.P.: Sfm with mrfs: Discrete-continuous optimization for large-scale structure from motion. IEEE TPAMI **35**(12), 2841–2853 (2012)
10. Cui, H., Gao, X., Shen, S., Hu, Z.: Hsfm: Hybrid structure-from-motion. In: CVPR. pp. 1212–1221 (2017)
11. Ding, Y., Li, K., Mei, H., Liu, S., Hou, G.: Watermono: Teacher-guided anomaly masking and enhancement boosting for robust underwater self-supervised monocular depth estimation. IEEE TIM (2025)
12. Ebner, L., Billings, G., Williams, S.: Metrically scaled monocular depth estimation through sparse priors for underwater robots. In: ICRA. pp. 3751–3757 (2024)
13. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
14. Huang, S., Wang, K., Liu, H., Chen, J., Li, Y.: Contrastive semi-supervised learning for underwater image restoration via reliable bank. In: CVPR. pp. 18145–18155 (2023)
15. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: CVPR. pp. 406–413 (2014)
16. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. In: 3DV (2026)
17. Kim, A., Eustice, R.: Pose-graph visual slam with geometric model selection for autonomous underwater ship hull inspection. In: IROS. pp. 1559–1565 (2009)
18. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: ECCV. pp. 71–91. Springer (2024)
19. Levy, D., Peleg, A., Pearl, N., Rosenbaum, D., Akkaynak, D., Korman, S., Treibitz, T.: Seathru-nerf: Neural radiance fields in scattering media. In: CVPR. pp. 56–65 (2023)
20. Li, H., Song, W., Xu, T., Elsig, A., Kulhanek, J.: Watersplatting: Fast underwater 3d scene reconstruction using gaussian splatting. In: 3DV. pp. 969–978 (2025)

21. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. In: ICLR (2026)
22. Lin, H., Liang, D., Qi, Z., Bai, X.: A unified image-dense annotation generation model for underwater scenes. In: CVPR. pp. 961–970 (2025)
23. Liu, R., Fan, S., Wang, W., Yang, Y.: Underwater visual slam with depth uncertainty and medium modeling. In: ICCV. pp. 970–980 (2025)
24. Liu, Y., Jiang, Q., Li, X., Luo, T., Ren, W.: Toward better than pseudo-reference in underwater image enhancement. IEEE TIP (2025)
25. Lv, Q., Dong, J., Li, Y., Chen, S., Yu, H., Zhang, S., Wang, W.: Uwstereo: A large synthetic dataset for underwater stereo matching. IEEE TCSVT (2025)
26. Maggio, D., Lim, H., Carlone, L.: VGGT-SLAM: Dense rgb slam optimized on the sl (4) manifold. In: NeurIPS (2025)
27. Randall, Y.: Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets. Master's thesis, University of Haifa (Israel) (2023)
28. Ren, J., Jiang, X., Li, Z., Liang, D., Zhou, X., Bai, X.: Minima: Modality invariant image matching. In: CVPR. pp. 23059–23068 (2025)
29. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR. pp. 4104–4113 (2016)
30. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. ACM TOG pp. 835–846 (2006)
31. Tang, Y., Zhu, C., Wan, R., Xu, C., Shi, B.: Neural underwater scene representation. In: CVPR. pp. 11780–11789 (2024)
32. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS. vol. 30 (2017)
33. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE TPAMI (1991)
34. Wang, C., Xu, H., Jiang, G., Yu, M., Luo, T., Chen, Y.: Underwater monocular depth estimation based on physical-guided transformer. IEEE TGRS **62**, 1–16 (2024)
35. Wang, H., Anantrasirichai, N., Zhang, F., Bull, D.: Uw-gs: Distractor-aware 3d gaussian splatting for enhanced underwater scene reconstruction. In: WACV. pp. 3280–3289 (2025)
36. Wang, H., Agapito, L.: 3d reconstruction with spatial memory. In: 3DV. pp. 78–89 (2025)
37. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR. pp. 5294–5306 (2025)
38. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR. pp. 20697–20709 (2024)
39. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. In: ICLR (2026)
40. Wu, Z., Wang, Y., Wen, Y., Zhang, Z., Wu, B., Tang, H.: Stereoadapter: Adapting stereo depth estimation to underwater scenes. In: ICRA (2026)
41. Xie, Y., Kong, L., Chen, K., Zheng, Z., Yu, X., Yu, Z., Zheng, B.: Uveb: A large-scale benchmark and baseline towards real-world underwater video enhancement. In: CVPR. pp. 22358–22367 (2024)
42. Xu, W., Wang, C., Liang, D., Zhao, Z., Jiang, X., Zhang, P., Bai, X.: Nautilus: A large multimodal model for underwater scene understanding. In: NeurIPS (2025)
43. Yang, D., Leonard, J.J., Girdhar, Y.: Seasplat: Representing underwater scenes with 3d gaussian splatting and a physically grounded image formation model. In: ICRA. pp. 7632–7638 (2025)

44. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: CVPR. pp. 21924–21935 (2025)
45. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: NeurIPS. vol. 37, pp. 21875–21911 (2024)
46. Yu, B., Wu, J., Islam, M.J.: Udepth: Fast monocular depth estimation for visually-guided underwater robots. In: ICRA (2023)
47. Zhang, F., You, S., Li, Y., Fu, Y.: Atlantis: Enabling underwater depth estimation with stable diffusion. In: CVPR. pp. 11852–11861 (2024)
48. Zhao, Z., Yang, L., Long, S., Pi, J., Zhou, L., Wang, J.: Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In: CVPR. pp. 11350–11359 (2023)
49. Zhou, J., Liang, T., Zhang, D., Liu, S., Wang, J., Wu, E.Q.: Waterhe-nerf: Water-ray matching neural radiance fields for underwater scene reconstruction. Information Fusion **115**, 102770 (2025)