AquaStereo: Enabling Underwater Stereo Matching via Depth-Conditioned Diffusion and Geometry Self-Distillation

Qizhe Wei*[1], Yingping Liang*[1], Shaodi You[2], and Ying Fu†[1]

[1] Beijing Institute of Technology, Beijing, China
[2] University of Amsterdam, Amsterdam, The Netherlands
{qizhewei,liangyingping,fuying}@bit.edu.cn s.you@uva.nl

Abstract. Learning-based stereo matching models struggle in underwater environments due to scarce in-domain data and the difficulty of extracting discriminative correspondences from degraded imagery. In this work, we present **AquaStereo**, a perception-enhanced framework with a data simulation pipeline and a self-distillation strategy that jointly address data scarcity and feature degradation in underwater stereo matching. First, a depth-conditioned diffusion pipeline renders underwater stereo pairs while preserving binocular geometry, with a lightweight left-right consistency module ensuring geometric alignment. Training on this synthetic corpus effectively narrows the terrestrial–underwater gap and improves zero-shot robustness. Second, a frozen binocular teacher trained on clean terrestrial pairs guides a student exposed to rendered underwater pairs with perturbations. A stage-weighted sequence loss is performed to align the student's disparities with the teacher's geometry, while a clean-branch supervision with shared pseudo targets prevents scale drift. To further enhance feature stability under turbidity and low texture, we introduce learnable perception frames, a perception-enhanced feature formulation that constructs robust matching descriptors by fusing temporal cues from two auxiliary views encoded by a video backbone with semantic features extracted by a strong image encoder. Extensive experiments demonstrate that **AquaStereo** substantially improves robustness and zero-shot generalization in challenging underwater scenarios.

Keywords: Stereo matching · Diffusion model · Underwater vision

1 Introduction

Stereo matching is a fundamental task in computer vision, spanning autonomous driving, robotics, 3D reconstruction, and virtual reality [18,38,56]. Accurate underwater depth is likewise critical for autonomous underwater vehicles and marine applications [3,7,32,40,55]. Despite remarkable progress in terrestrial stereo matching [5, 12, 13, 27], direct transfer to underwater scenes remains ineffective

* Equal contribution.
† Corresponding author.

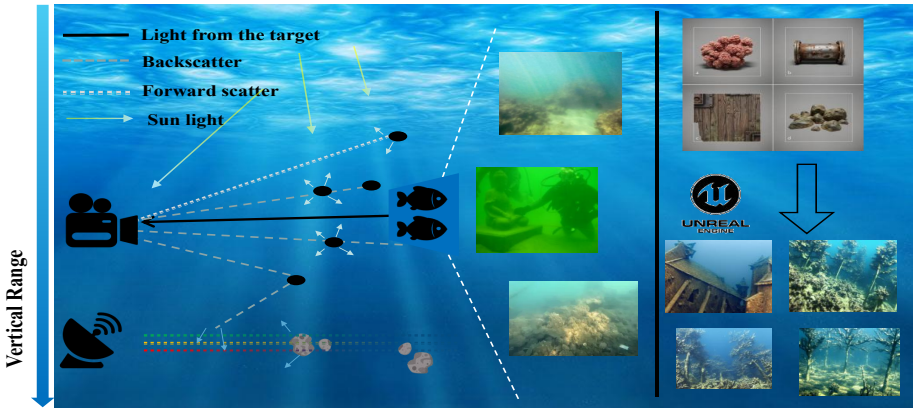


Fig. 1: Principles of underwater data acquisition. Left: Underwater sensing with LiDAR and vision cameras. **Right:** CG-based underwater image rendering.

because underwater image formation differs fundamentally from terrestrial settings. Attenuation, scattering, and backscatter reduce contrast and distort color, violating common photometric assumptions and suppressing reliable textures. More critically for stereo, these degradations are often view-dependent, breaking left–right consistency and making cost-volume matching ambiguous. A key bottleneck is data. As illustrated in Figure 1, underwater stereo data are typically obtained either by real-world capture or by CG rendering, yet both have clear limitations. Real capture is expensive and hard to scale; moreover, scattering and refraction degrade both passive vision and active ranging (e.g., LiDAR), often resulting in sparse or unreliable depth/disparity supervision (e.g., HIMB [42], FLSea [34], Squid [4]). In contrast, CG rendering/simulation provides dense supervision at low cost (e.g., VAROS [65], UWStereo [28]), but lacks photorealism and introduces a noticeable domain gap to real underwater imagery.

To tackle the challenge of scarce and unreliable underwater stereo data, we present **AquaStereo**. We take a third route: generation. Compared with real capture and CG rendering, diffusion-based generation [43, 62] can produce more realistic underwater appearances while remaining highly controllable and scalable. However, underwater stereo generation is more challenging than single-image translation [16, 60] or generic image generation [35]: it must preserve left–right correspondence and epipolar consistency under complex and highly variable underwater degradations. Violating this structure directly corrupts cost-volume matching in modern stereo pipelines [5, 53]. To control these non-deterministic appearance changes, we build a physics-inspired prompt pool encoding water types and degradation cues [2], constructed from CLIP-retrieved [33] underwater imagery and LLM-curated underwater descriptions. We further incorporate a lightweight left–right consistency module to enforce binocular alignment, ensuring generated pairs remain suitable for downstream stereo training.

Building on the generated data, we further improve robustness via cross-domain self-distillation: a frozen teacher provides stable geometric supervision, while the student learns from underwater pairs with perturbations; clean-branch supervision with shared pseudo targets prevents scale drift [44, 61]. Finally, we introduce a perception-enhanced matcher to stabilize features under turbidity and low-texture conditions by concatenating two learnable auxiliary views with the stereo pair, encoding them with a video backbone, and fusing image-level semantics from an image encoder to produce robust matching descriptors [31,63].

Extensive experiments demonstrate that our pipeline significantly boosts underwater stereo depth estimation accuracy. In summary, our contributions are: (1) we propose **AquaStereo**, an underwater stereo matching framework that improves generalization and accuracy without target-domain fine-tuning; (2) we introduce a depth-conditioned diffusion pipeline with a physics-inspired prompt pool and a lightweight left–right consistency module for stereo generation; (3) we design a cross-domain self-distillation strategy and a perception-enhanced matcher that improve robustness on underwater degradations.

2 Related Work

Stereo Matching Datasets. Learning-based stereo matching primarily relies on annotated terrestrial datasets for supervised training, while underwater resources such as HIMB [42], FLSea [34], VAROS [65], and UWStereo [28] remain limited in different ways, as summarized in Figure 2. Some offer realistic imagery but lack dense ground truth depth or disparity (e.g., only relative depth or sparse labels). In contrast, others are synthetic with simplified optics, reduced photorealism, or restricted scene diversity and resolution, largely due to the cost and difficulty of underwater acquisition. To alleviate data scarcity, synthetic pipelines have been explored, including game-engine rendering, video forwarding to derive pseudo pairs, and 2D affine augmentations; although these strategies increase volume, they fail to reproduce underwater attenuation, backscatter, and turbidity, leaving a clear domain gap in real data. In contrast, our approach generates large-scale, diverse training sets from real-world stereo images by rendering geometry-preserving underwater counterparts, narrowing this gap without additional underwater collection.

Image Editing Models. Latent diffusion models [35, 57, 59, 64] generate high-fidelity images in a learned latent space and provide stronger stability and controllability than GAN-based stylization [8, 10], especially when combined with classifier-free guidance [19] and ControlNet-style structural conditioning [36, 62]. Beyond image synthesis, diffusion priors have also been used for dense prediction tasks such as optical flow and monocular depth estimation [20, 23, 25, 26, 37]. Recent diffusion-based monocular depth methods further exploit this property by synthesizing challenging yet depth-consistent appearances from clean images, thereby improving robustness under adverse conditions [24, 45, 49, 58]. However, extending such image editing strategies from monocular depth to stereo matching is non-trivial: the generated left and right views must not only be individually

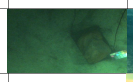
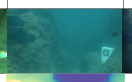
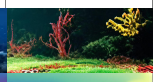
Datasets	HIMB	FLSea	VAROS	Squid	StereoAdapter	Ours
Image						
Disparity map	----		----			
Disparity	No	Estimated depth	No	Dense	Dense	Dense
Type	----	Relative depth	----	Metric depth	Metric depth	Metric depth
Acquisition	Real capture	Real capture	Blender rendered	Real capture	UE rendered	Diffusion rendered
Cost	Expensive	Expensive	Low	Expensive	Low	Low
Photorealism	Realistic	Realistic	Unrealistic	Realistic	Synthetic	Scene Realistic
Controllable Diversity	----	----	Difficult	----	Difficult	Easy
Image number	4062	7337	4713	57	40000+	42631 or more

Fig. 2: Statistics and comparison of existing underwater stereo datasets. Including HIMB [42], FLSea [34], VAROS [65], Squid [4], StereoAdapter [51] and our proposed dataset. The table summarizes key statistics such as the number of stereo pairs, capture conditions, and the availability of geometry annotations.

realistic, but also preserve cross-view correspondence, epipolar consistency, and disparity scale. Independently editing the two views may introduce inconsistent textures, boundaries, color shifts, or scattering patterns, which directly corrupt cost-volume matching.

Stereo Matching Networks. Learning-based stereo matching has replaced traditional pipelines with convolutional neural networks (CNNs). GCNet [21] first regularized the 4D cost volume using 3D convolutions. Subsequent models such as PSMNet [5], GANet [60] and GwcNet [15] improved accuracy via hierarchical aggregation, group-wise correlation, and guided aggregation, while the cascade-based CFNet [39] further enhanced efficiency and generalization. Building on these advances, IGEV [52] introduced a geometry encoding volume that captures non-local context. The field has since shifted toward more generalizable architectures: IGEV++ [53] advances geometry-aware encoding and iterative refinement, and vision foundation models [48] such as FoundationStereo [50] leverage large-scale pretraining to provide transferable backbones that set new baselines. However, most of these methods are developed for normal viewing conditions and degrade under underwater artifacts, making robustness in such challenging environments a key open problem.

3 Method

In this section, we first present the problem formulation and motivation. We then describe the data generation pipeline with self-distillation, shown in Fig-

ure 3. Finally, we detail the improved stereo matching network with a perception-enhanced module, shown in Figure 4.

3.1 Formulation and Motivation

Accurate zero-shot stereo matching underwater remains challenging: attenuation, backscatter, turbidity, wavelength dependent color casts, and low illumination reduce contrast and texture, making correspondence unreliable (Figure 1). A major bottleneck is data. Acquiring underwater stereo with accurate metric depth is costly and difficult—active sensors (e.g., underwater LiDAR/structured light) are expensive, range-limited, and often sparse, while stereo capture requires careful calibration under refraction and stable hardware in dynamic conditions. Consequently, real datasets (e.g., Sea-thru [2] and Squid [4]) remain small and supervision is limited. To mitigate data scarcity, prior work uses GAN-based [16,22] translation and style transfer to make terrestrial images look "underwater," but results are often stylized and still exhibit a clear domain gap.

This is where **AquaStereo** comes into play. We introduce a scalable depth-conditioned generation pipeline that synthesizes diverse underwater stereo pairs from depth maps and concise textual prompts, enabling virtually unlimited sampling while keeping geometric structure intact. Building on this data, we further employ cross-domain self-distillation to transfer reliable geometric cues from clean-domain stereo to underwater statistics. Finally, we augment the matcher with learnable perception frames that absorb underwater priors during training, improving feature extraction under severe degradations. We also incorporate a lightweight left-right consistency module to reduce stereo inconsistencies and improve geometric coherence.

3.2 Stereo Data Generation for Underwater Scenes

To address the scarcity of underwater stereo data, we build a scalable simulation pipeline that converts terrestrial stereo pairs into geometry-preserving underwater counterparts. Since dense, reliable disparity supervision is often sparse or noisy in real-world terrestrial datasets, we first collect a large corpus of high-quality rectified stereo pairs and compute pseudo disparities using a strong foundation stereo model (e.g., FoundationStereo [50]):

$$D_{\mathrm{GT}} = F_{FStereo}(I_L, I_R). \tag{1}$$

Each sample $(I_L, I_R, D_{\mathrm{GT}})$ lies in the normal-domain dataset. Our goal is to map it to the underwater domain with an operator that changes appearance while preserving binocular geometry.

Depth-conditioned diffusion generation. Underwater degradation depends strongly on scene depth. We synthesize underwater appearance by conditioning a diffusion model on per-pixel depth maps [54] $(Depth_L, Depth_R)$, enabling distance-dependent haze and scattering without simply copying terrestrial colors. As illustrated in Figure 3 (prompt generation), we construct a diverse text

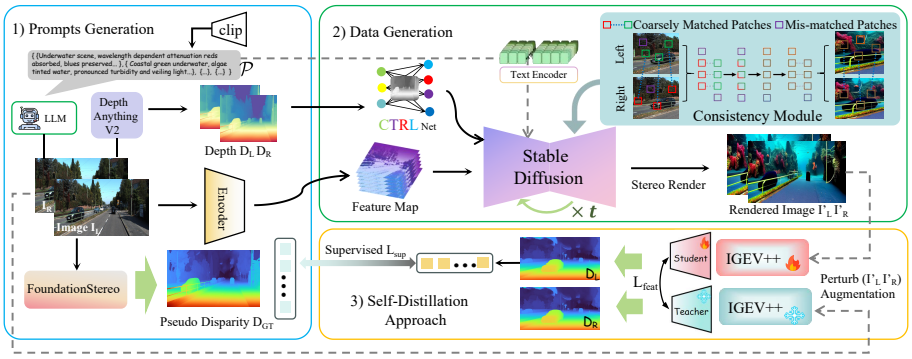


Fig. 3: Overview of our framework. The pipeline comprises three parts: (1) Prompts Generation: brief underwater descriptors (via an LLM or CLIP [33]) and a monocular depth map is extracted from a clean terrestrial pair to describe underwater scenes. (2) Data Generation: the text prompt and depth map condition a ControlNet-guided diffusion model; a lightweight patch-level consistency module enforces left-right coherence and yields geometry-preserving underwater stereo images. (3) Self-Distillation Approach: the frozen teacher operates on the clean pair, while the student ingests the rendered/perturbed underwater pair.

prompt pool $\mathcal{P}$ by: (i) collecting underwater-scene descriptions and extracting representative prompts via CLIP [33]; and (ii) using an LLM to summarize key principles of underwater image formation from physics-based literature into concise prompts (e.g., wavelength-dependent attenuation, backscatter, and veiling light [1, 2]). During generation, we randomly sample a prompt $p \sim \mathcal{P}$ and feed the depth maps together with p into a ControlNet to obtain the guidance features c_t (Figure 3, data generation). Conditioned on c_t, we use Stable Diffusion to synthesize the underwater stereo pair (I'_L, I'_R):

$$c_t = F_{\text{CtrlNet}}(z_t, Depth_L, Depth_R, p), \tag{2}$$

$$(I'_L, I'_R) = F_{\text{StableDiffusion}}(z_t, p \mid c_t). \tag{3}$$

This procedure yields diverse yet geometry-faithful underwater stereo pairs.

Coherence-Enhanced Consistency. Unlike single-image generation, diffusion may introduce stereo pair inconsistencies. We mitigate this using a lightweight module that samples patch pairs along epipolar lines, scores them with appearance cues and coarse disparity priors, and applies a single self-attention pass over mixed positive and negative matches. Concretely, for a left patch feature $\mathbf{f}_i^L$ and candidate right features $\mathbf{f}_j^R$, we define a disparity-aware similarity:

$$s_{ij} = \lambda \frac{\langle \mathbf{f}_i^L, \mathbf{f}_j^R \rangle}{\|\mathbf{f}_i^L\|_2 \, \|\mathbf{f}_j^R\|_2} + (1-\lambda) \exp\Big(- \tfrac{|d_i - d_j|}{\tau} \Big), \tag{4}$$

where d_i and d_j are coarse disparity priors, and $\lambda \in [0, 1]$ balances appearance and geometric cues. The resulting scores are normalized into soft confidence

masks, which modulate the ControlNet features and improve binocular coherence with negligible overhead. The geometry-consistent synthetic pairs form our underwater training dataset. We summarize our dataset alongside representative underwater benchmarks in Figure 2. Compared with real capture (costly and hard to scale) and CG rendering (low cost but often less photorealistic and harder to diversify), our diffusion-based generation provides a low-cost and scalable alternative with scene-realistic appearance and controllable diversity.

3.3 Self-Distillation with Geometry Alignment

To more tightly couple terrestrial and underwater domains and extend effective supervision across both, we adopt cross-domain self-distillation. For each geometry-sharing pair $\mathbf{x}^{\text{norm}} = (I_L, I_R, D_{GT})$ and $\mathbf{x}^{\text{underwater}} = (I'_L, I'_R, D_{GT})$, we employ a teacher–student scheme tied together by the common target D_{GT}. The teacher T is trained on clean terrestrial pairs and remains frozen throughout distillation. To further enhance the student's learning from the teacher model, increase the diversity of underwater appearances, and improve robustness to underwater domain shifts, we apply an additional augmentation to the underwater branch:

$$(\tilde{I}'_L, \tilde{I}'_R) = \text{Perturb}(I'_L, I'_R), \tag{5}$$

where Perturb$(\cdot)$ includes turbidity/backscatter amplification, color shifts, compression noise, and mixed or mismatched patch augmentations.

Self-Distillation Approach. As shown in Figure 3 (Self-Distillation Approach), given the terrestrial pair (I_L, I_R), the teacher extracts multi-scale features F_T. For each underwater counterpart $(\tilde{I}'_L, \tilde{I}'_R)$, the student extracts features F_S. We align geometry across domains by matching features at the feature-extraction stage. Concretely, we apply an ℓ_1 feature alignment loss with a stage weight:

$$\mathcal{L}_{\text{feat}} = \sum_{k=1}^{K} \left\| F_S(\tilde{I}'_L, \tilde{I}'_R) - F_T(I_L, I_R) \right\|_1, \tag{6}$$

where $F_T(\cdot)$ and $F_S(\cdot)$ denote the teacher and student feature extractors, respectively, and K is the number of perturbed underwater counterparts sampled via Perturb$(\cdot)$ for the same geometry-sharing pair. This feature alignment encourages the student to retain teacher-consistent geometric cues under underwater degradations without requiring any weight sharing.

Supervised loss for scale stability. To prevent source-domain drift and preserve metric scale, we supervise the student on the unperturbed pair using the shared pseudo disparity D_{GT}:

$$\mathcal{L}_{\text{sup}} = \sum_{k=1}^{K} \left\| D_S(I'_L, I'_R) - D_{\text{GT}} \right\|_1, \tag{7}$$

where $D_S(\cdot)$ denotes the student's disparity prediction. This term stabilizes iterations and calibrates the student to the teacher before transferring to underwater

statistics. In sum, the training objective includes the two sequence terms above:

$$\mathcal{L} = \mathcal{L}_{\text{feat}} \; + \; \lambda_{\text{sup}} \, \mathcal{L}_{\text{sup}}, \tag{8}$$

with no additional regularization. We train jointly on both domains, use identical parameters for the two branches, keep the teacher frozen and parameter-disjoint, and apply perturbations only to the underwater inputs so as to widen the non-semantic gap while still preserving geometry through the shared target D_{GT}.

3.4 Perception-Enhanced Network

Learning a perception-enhanced encoder. Underwater stereo matching is highly sensitive to feature degradation. Attenuation, backscatter, turbidity, and low texture weaken local correspondences and make cost-volume construction ambiguous, especially around object boundaries and texture-poor regions. Standard image encoders usually extract features from the left and right views, which limits their ability to model cross-view interactions before matching. Although a rectified stereo pair is not a temporal video, a video encoder provides a natural mechanism for cross-frame attention. We therefore formulate the stereo pair as a pseudo temporal cross-view sequence, where cross-frame attention serves as a proxy for cross-view feature alignment. To enrich this pseudo temporal context and capture underwater degradation priors, we augment a stereo pair (I_L, I_R) with two learnable perception frames I_{P1} and I_{P2}, as shown in Figure 4. The four frames are arranged in temporal order and passed through a video encoder to obtain multi-scale features:

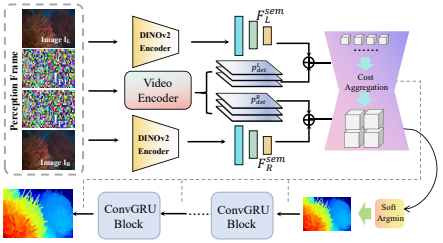


Fig. 4: Perception-enhanced stereo matcher. We append two perception frames to the stereo pair, fuse the encoded features with DINOv2 [31] semantics, and iteratively refine multi-range cost volumes to predict disparity.

$$\mathbf{f}_{left}, \mathbf{f}_{right} \; = \; \mathcal{F}_{\text{enc}}\big(I_{P1} \; \textcircled{c} \; I_L \; \textcircled{c} \; I_R \; \textcircled{c} \; I_{P2}\big), \tag{9}$$

where $\textcircled{c}$ denotes temporal concatenation and $\mathcal{F}$ is a video encoder following [63]. Let f_{left} and f_{right} denote the view-aligned features extracted from the left and right frames, respectively, via temporal modeling. The perception frames are learnable parameters that act as degradation-aware visual prompts. During training, they absorb underwater priors and help the encoder produce more stable, view-consistent, and discriminative features under challenging degradations. In parallel, we extract high-level semantics from the raw stereo views using a DINOv2 [31] image backbone: $F_L^{\text{sem}} = \mathcal{E}_{\text{img}}(I_L)$ and $F_R^{\text{sem}} = \mathcal{E}_{\text{img}}(I_R)$, where $\mathcal{E}_{\text{img}}$ is a shared-weight feature extractor. We then fuse the perception features

Table 1: Comparison with state-of-the-art stereo methods on the UW-Stereo [28] benchmark. "Ours" is trained on the rendered underwater dataset, whereas most baselines are trained on SceneFlow [29], except LightStereo [14], StereoAnything [13], NMRF [11], MonSter [6], and FoundationStereo [50], which use mixed synthetic and real data.

Method	Coral		Default		Industry		Ship		Total	
	EPE↓	D1↓	EPE↓	D1↓	EPE↓	D1↓	EPE↓	D1↓	EPE↓	D1↓
PSMNet [5]	2.680	13.170	2.360	7.210	2.850	10.580	4.950	15.700	3.540	12.360
CFNet [39]	2.503	10.043	2.060	7.615	2.605	10.518	4.903	16.123	3.368	12.092
GwcNet [15]	2.600	10.668	4.162	9.750	4.101	13.367	7.438	20.295	5.138	14.986
IGEV [52]	2.262	12.310	1.088	5.580	1.990	7.590	2.654	12.300	2.138	9.758
LightStereo [14]	2.503	11.209	1.776	7.542	2.551	10.598	3.966	16.668	2.952	12.487
StereoAnything [13]	2.117	8.518	1.118	5.539	1.527	8.058	3.087	13.187	2.138	9.655
NMRF [11]	3.036	14.580	2.734	10.570	2.676	12.610	5.353	18.390	3.747	14.758
MonSter [6]	1.882	9.010	<u>0.664</u>	<u>2.940</u>	1.061	4.870	<u>1.254</u>	<u>6.340</u>	<u>1.195</u>	<u>5.744</u>
FoundationStereo [50]	<u>1.811</u>	<u>7.590</u>	0.844	3.090	<u>0.718</u>	<u>2.490</u>	2.750	8.890	1.669	5.768
AquaStereo (Ours)	**1.528**	**7.180**	**0.338**	**1.320**	**0.373**	**1.510**	**0.500**	**2.240**	**0.590**	**2.616**

with their semantic counterparts via channel concatenation followed by a 1×1 projection, yielding the left-right matching descriptors:

$$p_L \;=\; g([\, f_{\text{left}};\, F_L^{\text{sem}}\,]), \quad p_R \;=\; g([\, f_{\text{right}};\, F_R^{\text{sem}}\,]). \tag{10}$$

The fused features (p_L, p_R) are used to build the cost volume and are fed into an iterative refinement with Soft-Argmin regression, following IGEV++ [53].

4 Experiments

In this section, we first introduce the datasets and evaluation metrics for experiments. Then, detailed comparisons are conducted with the state-of-the-art methods. Finally, ablations and discussions are performed to confirm the effectiveness of the main components.

4.1 Experimental Setup

Datasets. We use terrestrial stereo datasets for supervised training and underwater benchmarks for evaluation. For training, SceneFlow [29], KITTI 2012 [9], and KITTI 2015 [30] provide clean stereo pairs with reliable geometry to drive our rendering and learning pipeline. For underwater evaluation, we report results on FLSea-Stereo [34] and Squid [4] as real in-situ, forward-looking benchmarks captured under diverse underwater conditions. We additionally evaluate on UW-Stereo [28] as a physics-based synthetic benchmark with dense metric ground truth, and on the underwater split of TartanAir [47] to assess generalization to simulated underwater scenes.

Stereo matching models. We compare our approach with representative stereo matchers: PSMNet [5], CFNet [39], GwcNet [15], IGEV [52], IGEV++ [53],

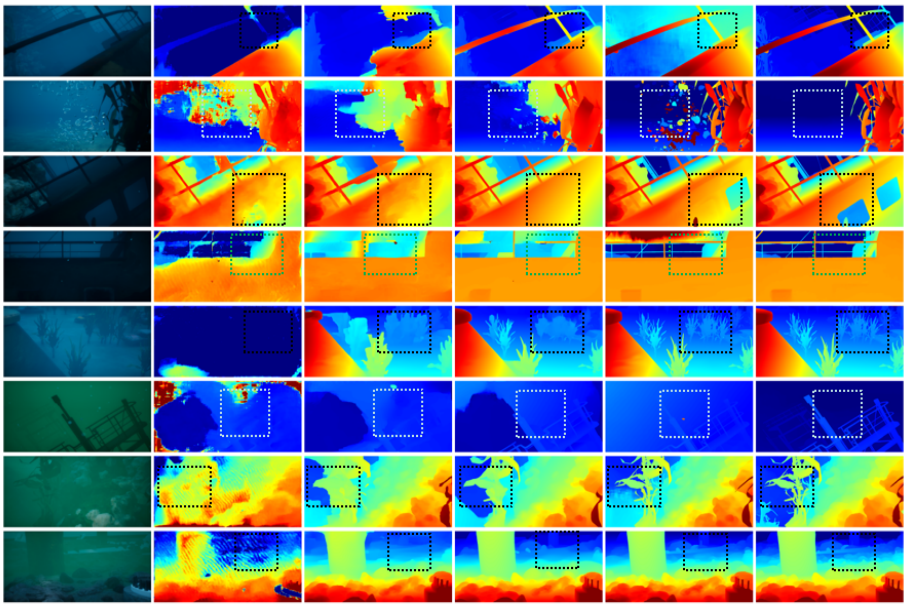


Fig. 5: Qualitative results in underwater scenes. Our method produces more coherent disparity maps under underwater degradations, preserving sharper boundaries and finer structures with fewer artifacts.

LightStereo [14], StereoAnything [13], NMRF [11], MonSter [6], and FoundationStereo [50]. For the perception-enhanced encoder, we additionally use a Change3D-style video encoder [63] and DINOv2 [31]. For self-distillation, both the teacher and student adopt the same IGEV++ architecture; the teacher is pre-trained on our terrestrial dataset using pseudo-disparities generated by FoundationStereo, and is then used to supervise the student during training. Unless otherwise stated, we follow official implementations and default settings.

Stable Diffusion models. We implement our generation pipeline with the Diffusers library, adapting Stable Diffusion (SD) and ControlNet. We use SD v1.5 as the image generator with a DDIM scheduler and 50 sampling steps. For depth-conditioned generation, we employ ControlNet (control_v11f1p_sd15_depth) together with Depth-Anything V2 for monocular depth estimation. Source images from SceneFlow and KITTI are used to synthesize stereo data in underwater conditions. We refer to the resulting synthetic training set as UW-Dataset (40K).

Training Parameters. Unless otherwise stated, we train with a learning rate of 10^{-4} (batch size = 8 per GPU). Models are optimized at a resolution of 384×736 pixels, with the maximum disparity kept fixed across methods. Training takes approximately three days on 4 RTX 4090 (24GB) GPUs. For self-distillation, we adopt the frozen-teacher and student losses described in Sec. 3.3, use the same model configuration, and set $\lambda_{\text{sup}} = 0.2$.

Table 2: Comparison on UWStereo with different training sets. We evaluate three representative stereo networks trained on SceneFlow [29], KITTI [30], or our rendered UW-Dataset.

Datasets	Networks	Total	
		EPE↓	D1↓
SceneFlow [29]	PSMNet [5]	3.5400	12.360
KITTI [30]		4.7630	17.558
UW-Dataset		**1.1422**	**10.16**
SceneFlow [29]	StereoBase [12]	2.5930	10.861
KITTI [30]		3.8360	14.755
UW-Dataset		**2.5070**	**10.402**
SceneFlow [29]	IGEV++ [53]	2.1380	9.758
KITTI [30]		4.0100	16.106
UW-Dataset		**1.7280**	**8.418**

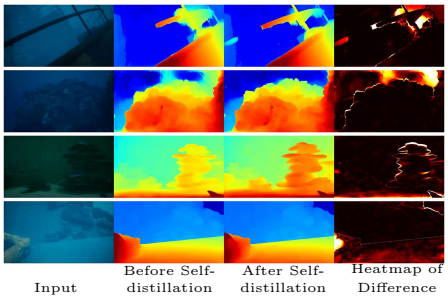


Fig. 6: Effect of self-distillation on disparity maps. Self-distillation produces smoother and more complete disparities, while the heatmap highlights corrections around structural boundaries and degraded regions.

Evaluation Metrics. Zero-shot generalization is evaluated on UWStereo [28], FLSea [34], Squid [4] and TartanAir [47]. We report EPE and the D1 outlier rate as our primary evaluation metrics. EPE is defined as the mean absolute disparity error (in pixels) over all valid pixels. D1 is the percentage of pixels with error larger than $\max(3\text{ px}, 0.05\,D)$, which serves as a stringent indicator of robustness and stability in challenging underwater scenes.

4.2 Quantitative Results

Comparison with stereo matching methods. Table 1 compares stereo matching methods on UWStereo with four scene splits: Coral, Default, Industry, and Ship. Our model ranks first on all splits and overall, achieving the lowest Total EPE↓/D1↓ of 0.59/2.61. Notably, methods that rely heavily on photometric or brightness-constancy assumptions, such as NMRF [11], degrade underwater, where absorption, scattering, and backscatter violate these priors. Moreover, although many baselines are trained on mixed datasets, our approach remains strong, highlighting the effectiveness of our pipeline, which combines depth-conditioned synthesis, a lightweight left–right consistency check, dual-domain self-distillation, and a perception-enhanced stereo framework.

Comparison with other training sets. Table 2 and Table 3 compare models trained on SceneFlow [29], KITTI [30], and our rendered UW-Dataset and evaluated zero-shot on the four UWStereo splits; training on UW-Dataset yields the best total accuracy for all three representative architectures (PSMNet [5], StereoBase [12], IGEV++ [53]) with consistent gains, and IGEV++ represents the strongest model. More detailed results for all tables are provided in the supplementary material. Note that models trained only on terrestrial data (SceneFlow/KITTI) degrade notably underwater, reflecting the domain gap caused by absorption, scattering, and backscatter. Also, by fixing the architecture to

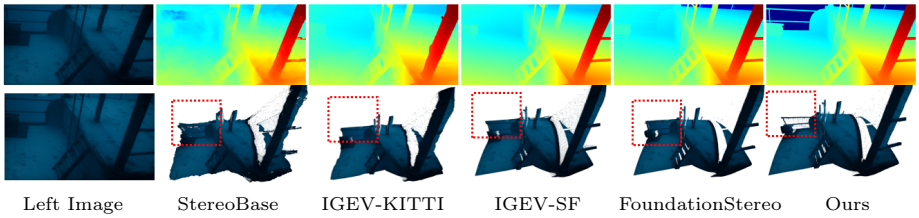


Fig. 7: Qualitative point-cloud reconstructions in underwater scenes. Compared with multiple baselines, our method produces denser and cleaner reconstructions with fewer holes, floating outliers, and depth artifacts.

Table 3: Comparison on underwater benchmarks with different training sets. We evaluate IGEV++ [53] trained on SceneFlow [29], KITTI [30], UWStereo [28], StereoAdapter [51] or our rendered UW-Dataset on three underwater datasets: FLSea (Canyon/Rock) [34], Squid [4], and TartanAir [47].

Datasets	Networks	FL-Canyon [34]		FL-Rock [34]		Squid [4]		TartanAir [47]	
		EPE↓	D1↓	EPE↓	D1↓	EPE↓	D1↓	EPE↓	D1↓
SceneFlow [29]	IGEV++ [53]	5.41	35.82	5.09	22.62	2.18	12.28	0.62	7.81
SF [29]+KITTI [30]		5.10	22.76	5.20	21.82	2.02	11.55	0.60	7.43
UWStereo [28]		5.01	21.45	4.96	20.73	2.08	12.64	0.63	7.85
StereoAdapter [51]		<u>4.05</u>	<u>20.51</u>	<u>4.15</u>	<u>19.58</u>	<u>1.86</u>	<u>10.98</u>	**0.52**	**6.81**
UW-Dataset (Ours)		**3.05**	**18.31**	**3.13**	**17.44**	**1.80**	**10.52**	<u>0.55</u>	<u>6.92</u>

IGEV++ and varying only the training data, we compare several training sets of comparable scale, including SceneFlow, SceneFlow+KITTI, StereoAdapter [51], and our UW-Dataset (Table 3). This indicates that training on terrestrial data alone (SceneFlow or SceneFlow+KITTI) generalizes poorly to underwater benchmarks. StereoAdapter yields only limited gains on real underwater scenes, likely because CG rendering cannot fully reproduce real underwater image formation. In contrast, our UW-Dataset consistently achieves better performance across FLSea [34], Squid [4], and TartanAir [47], suggesting that our diffusion-based generation produces more realistic underwater appearances than CG rendering and leads to more effective training.

Qualitative Results. As shown in Figure 5, the disparity visualizations highlight the robustness of our method under strong color cast, depth-dependent attenuation, and heavy backscatter. In challenging regions such as the ship structure and surrounding rigging, our predictions maintain crisp object boundaries and smooth planar gradients, whereas baselines often exhibit bleeding artifacts and fragmented surfaces. Notably, transparent or semi-transparent phenomena (e.g., bubble-like structures and waterborne particles) frequently induce spurious matches in prior methods; in contrast, our results suppress these outliers and yield more coherent geometry. Thin structures and vegetation also remain continuous across the scene, preserving fine details in near- and mid-range areas that are critical for reliable underwater perception.

Table 4: LR-consistency ablation on UWStereo [28]. We evaluate LR consistency in terms of disparity quality and generation efficiency. Gen. Time is the average time per image, and Break px denotes LR-inconsistent pixels per 100 pixels.

Module	Methods	EPE↓	D1↓	Gen. Time (s) ↓	Break px↓
LR-Consistency Module	Off	3.237	14.82	8.4	15
	On	**2.953**	**12.14**	**8.1**	**5**

Moreover, Figure 7 shows that these improvements translate into higher-quality 3D reconstructions. The reconstructed point clouds from our disparities are denser and more complete, with fewer floating outliers and reduced depth discontinuity noise. Fine obstacle geometry becomes clearly visible: for example, fence-like structures and slender bars are recovered with consistent depth and sharp contours. Such obstacles can be missed when disparities are noisy or over-smoothed, potentially compromising downstream tasks (e.g., underwater robot navigation and inspection) where failing to detect barriers may lead to unsafe trajectories or collisions. We also observe more stable seafloor topology and more complete hull surfaces, indicating that improvements are not limited to a single object type but generalize across both textured and low-texture regions. Overall, the point-cloud evidence corroborates that our more accurate disparities enable faithful, structure-preserving underwater 3D reconstruction.

4.3 Ablation Study

Coherence-Enhanced Consistency Module. In this ablation, we isolate the effect of left-right (LR) consistency module under the same data-generation setting. As shown in Table 4, disabling the module during generation degrades the downstream stereo training results. Without enforcing cross-view coherence, the generated stereo pairs exhibit more inconsistencies, weakening correspondence cues and making matching ambiguous. Enabling our LR-consistency module reduces the inconsistency rate (Break px) from 15 to 5 with comparable generation cost, leading to improved EPE and D1.

Effectiveness of the self-distillation method. A binocular stereo matcher typically follows a three-stage pipeline: feature extraction, cost aggregation, and iterative refinement, as illustrated in Figure 4. Under identical training conditions, we evaluated several teacher–student coupling strategies by sharing weights at different stages of this pipeline. As reported in Table 5, parameter-disjoint self-distillation already improves robustness over the non-distilled baseline. Sharing the feature extractor, the iterative refinement, or the aggregation module between teacher and student yields comparable performance and only modest additional gains, suggesting that the benefit mainly comes from cross-domain supervision rather than architectural coupling. In contrast, adding ℓ_2 regularization restricts adaptation and degrades accuracy. The strongest performance is achieved when underwater-branch distillation is coupled with supervised learning on the clean branch using the shared pseudo target D_{GT}, underscoring the value of geometry-aligned cross-domain training.

Table 5: Ablation of self-distillation variants on the UWStereo benchmark [28]. All methods use the same IGEV++ [53] stereo backbone. We compare a baseline without self-distill to self-distill variants including without weight sharing, shared extractor, shared iterative refinement, shared cost aggregation, L2 regularization, and plus supervised.

Distillation variant	Total	
	EPE↓	D1↓
Baseline without self-distill	2.9530	12.144
Self-distill without weight sharing	2.7510	11.195
Self-distill + Share extractor	2.7740	11.262
Self-distill + Share iterative	2.8030	11.265
Self-distill + Share aggregation	<u>2.7200</u>	<u>11.120</u>
Self-distill + L2 regularization	3.0510	13.838
Self-distill + supervised	**2.6350**	**10.549**

Table 6: Ablation of feature designs on the UWStereo benchmark [28]. All variants share the same IGEV++ [53] stereo matcher and differ only in the feature extractor, including IGEV++ backbone, VGG19 [41]+DINOv2 [31], FoundationStereo features [50], VGGT-style attention block [46], ResNet [17]+DINOv2 [31], and AquaStereo (ours).

Feature extractor variant	Total	
	EPE↓	D1↓
IGEV++ backbone	<u>2.9530</u>	<u>12.144</u>
VGG19 + DINOv2	3.0790	12.955
FoundationStereo features	3.9290	15.429
VGGT-style attention	3.4090	15.462
ResNet + DINOv2	3.1410	13.197
AquaStereo (Ours)	**2.4830**	**10.824**

Effectiveness of the perception-enhanced encoder. Table 6 compares feature extractor designs under the same training protocol and IGEV++ stereo head, so performance differences can be attributed to the encoder. We evaluate several alternatives, including CNN backbones (VGG19 [41]/ResNet [17]) augmented with DINOv2 semantics [31], a VGGT-style attention block [46], and a FoundationStereo-like extractor [50]. While these variants improve representation capacity, they still struggle to produce stable correspondences under underwater degradations, where contrast loss, backscatter, and color casts weaken texture cues and amplify ambiguous matches. In contrast, **AquaStereo** achieves the best Total EPE/D1 (2.483/10.824), outperforming the IGEV++ backbone (2.953/12.144) and all other designs. We attribute the gains to two factors: learnable perception frames provide view-aligned, degradation-adaptive cues, while DINOv2 contributes high-level semantics that remain reliable when local texture is corrupted. Their fusion yields more discriminative matching descriptors, reducing cost-volume ambiguity and improving final disparities.

Visualizing self-distillation with heatmaps. As shown in Figure 6, we compare predictions before and after self-distillation and visualize their differences with heatmaps. The distilled model yields more coherent depth estimates, with sharper object boundaries, better-preserved thin structures, and fewer local artifacts such as flying pixels and halo effects. It also improves large planar regions by reducing depth bleeding and fragmentation in low-texture underwater areas. The heatmaps reveal that the main corrections are concentrated around depth discontinuities, specular regions, and haze-degraded areas, where stereo correspondence is typically ambiguous. These observations suggest that self-distillation enhances both local detail recovery and global depth consistency, leading to more reliable underwater geometry.

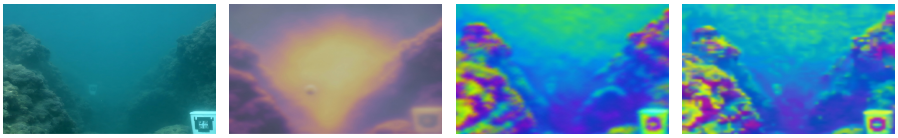


Fig. 8: Visualization of attention map and perception frame effects.

Visualizing perception frames. To better understand the role of perception frames, we visualize the attention responses of the stereo encoder with and without the proposed perception frames in Figure 8. Without perception frames, the attention is scattered over degraded regions, especially around low-texture areas, strong color casts, and backscatter-corrupted boundaries, leading to less stable correspondence cues. In contrast, after introducing perception frames, the model produces more concentrated and structure-aware attention on object contours and reliable matching regions. This suggests that the learnable perception frames enrich the pseudo temporal context, absorb underwater priors during training, and guide cross-frame attention toward view-consistent structures.

5 Conclusion

We present **AquaStereo**, a perception-enhanced framework that improves underwater stereo matching by jointly addressing data scarcity and feature degradation. On the data side, we introduce a depth-conditioned diffusion pipeline that synthesizes geometry-faithful underwater stereo pairs, guided by a physics-inspired prompt pool encoding water types and image-formation cues to control global appearance. A lightweight left–right consistency module further enforces binocular structure and epipolar alignment during generation, narrowing the terrestrial–underwater gap. On the training side, cross-domain self-distillation transfers clean geometric cues from a frozen teacher to a student trained on underwater pairs with perturbations, while clean-branch supervision with shared pseudo targets prevents scale drift. On the model side, we fuse learnable perception frames encoded by a video backbone with high-level semantics from a strong image encoder to produce robust matching descriptors in turbid, low-texture conditions. Extensive zero-shot evaluations on UWStereo, FLSea, Squid, and TartanAir, together with ablations, demonstrate consistent gains in accuracy and stability. Overall, AquaStereo offers a practical recipe for reliable underwater stereo and motivates future work on stronger physics-guided generation, online adaptation, and real-time efficiency.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62331006, 625B2026), and the Fundamental Research Funds for the Central Universities.

References

1. Akkaynak, D., Treibitz, T.: A revised underwater image formation model. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6723–6732 (2018)
2. Akkaynak, D., Treibitz, T.: Sea-thru: A method for removing water from underwater images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1682–1691 (2019)
3. Bailey, G.N., Flemming, N.C.: Archaeology of the continental shelf: marine resources, submerged landscapes and underwater archaeology. Quaternary Science Reviews **27**(23-24), 2153–2165 (2008)
4. Berman, D., Levy, D., Avidan, S., Treibitz, T.: Underwater single image color restoration using haze-lines and a new quantitative dataset. IEEE transactions on pattern analysis and machine intelligence **43**(8), 2822–2837 (2020)
5. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5410–5418 (2018)
6. Cheng, J., Liu, L., Xu, G., Wang, X., Zhang, Z., Deng, Y., Zang, J., Chen, Y., Cai, Z., Yang, X.: Monster: Marry monodepth to stereo unleashes power. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6273–6282 (2025)
7. Coleman, D.F., Newman, J.B., Ballard, R.D.: Design and implementation of advanced underwater imaging systems for deep sea marine archaeological surveys. In: OCEANS 2000 MTS/IEEE Conference and Exhibition. Conference Proceedings (Cat. No. 00CH37158). vol. 1, pp. 661–665. IEEE (2000)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
11. Guan, T., Wang, C., Liu, Y.H.: Neural markov random field for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5459–5469 (2024)
12. Guo, X., Zhang, C., Lu, J., Duan, Y., Wang, Y., Yang, T., Zhu, Z., Chen, L.: Openstereo: A comprehensive benchmark for stereo matching and strong baseline. arXiv preprint arXiv:2312.00343 (2023)
13. Guo, X., Zhang, C., Zhang, Y., Wang, R., Nie, D., Zheng, W., Poggi, M., Zhao, H., Ye, M., Zou, Q., et al.: Stereo anything: Unifying zero-shot stereo matching with large-scale mixed data. arXiv preprint arXiv:2411.14053 (2024)
14. Guo, X., Zhang, C., Zhang, Y., Zheng, W., Nie, D., Poggi, M., Chen, L.: Lightstereo: Channel boost is all you need for efficient 2d cost aggregation. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 8738–8744. IEEE (2025)
15. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3273–3282 (2019)

16. Hambarde, P., Murala, S., Dhall, A.: Uw-gan: Single-image depth estimation and image enhancement for underwater images. IEEE Transactions on Instrumentation and Measurement **70**, 1–12 (2021)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
18. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on pattern analysis and machine intelligence **30**(2), 328–341 (2008)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
20. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9492–9502 (2024)
21. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE international conference on computer vision. pp. 66–75 (2017)
22. Li, C., Anwar, S., Porikli, F.: Underwater scene prior inspired deep underwater image and video enhancement. Pattern recognition **98**, 107038 (2020)
23. Li, H., Fu, Y.: Fcdfusion: A fast, low color deviation method for fusing visible and infrared image pairs. Computational Visual Media **11**(1), 195–211 (2025)
24. Liang, Y., Fu, Y.: Relation-guided adversarial learning for data-free knowledge transfer. International Journal of Computer Vision **133**(5), 2868–2885 (2025)
25. Liang, Y., Fu, Y., Hu, Y., Shao, W., Liu, J., Zhang, D.: Flow-anything: Learning real-world optical flow estimation from large-scale single-view images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
26. Liang, Y., Hu, Y., Shao, W., Fu, Y.: Distilling monocular foundation model for fine-grained depth completion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22254–22265 (2025)
27. Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International conference on 3D vision (3DV). pp. 218–227. IEEE (2021)
28. Lv, Q., Dong, J., Li, Y., Chen, S., Yu, H., Zhang, S., Wang, W.: Uwstereo: A large synthetic dataset for underwater stereo matching. arXiv preprint arXiv:2409.01782 (2024)
29. Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: CVPR. pp. 4040–4048 (2016)
30. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR. pp. 3061–3070 (2015)
31. Oquab, M., Darcet, T., Moutakanni, T.H., Vo, H.T., Szafraniec, M., Cata, R., Shi, D., Lenc, K., Haziza, D., Harchaoui, Z., Mairal, J., Yu, X., Bojanowski, P., Joulin, A., Lefaudeux, B., Caron, M.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
32. Paull, L., Saeedi, S., Seto, M., Li, H.: Auv navigation and localization: A review. IEEE Journal of oceanic engineering **39**(1), 131–149 (2013)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
34. Randall, Y.: Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets. University of Haifa (Israel) (2023)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
36. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
37. Saxena, S., Herrmann, C., Hur, J., Kar, A., Norouzi, M., Sun, D., Fleet, D.J.: The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. Advances in Neural Information Processing Systems **36**, 39443–39469 (2023)
38. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision **47**(1), 7–42 (2002)
39. Shen, Z., Dai, Y., Rao, Z.: Cfnet: Cascade and fused cost volume for robust stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13906–13915 (2021)
40. Shortis, M., Abdo, E.H.D.: A review of underwater stereo-image measurement for marine biology and ecology applications. Oceanography and marine biology pp. 269–304 (2016)
41. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015), arXiv:1409.1556
42. Skinner, K.A., Zhang, J., Olson, E.A., Johnson-Roberson, M.: Uwstereonet: Unsupervised learning for depth estimation and color correction of underwater stereo imagery. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 7947–7954. IEEE (2019)
43. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
44. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
45. Tosi, F., Ramirez, P.Z., Poggi, M.: Diffusion models for monocular depth estimation: Overcoming challenging conditions. In: European Conference on Computer Vision. pp. 236–257. Springer (2024)
46. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
47. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4909–4916. IEEE (2020)
48. Wang, Y., Liang, Y., Fu, Y.: Boosting zero-shot stereo matching using large-scale mixed images sources in the real world. arXiv preprint arXiv:2505.08607 (2025)
49. Wang, Y., Liang, Y., Hu, Y., Fu, Y.: Robustereo: Robust zero-shot stereo matching under adverse weather. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25134–25144 (2025)

50. Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: Foundationstereo: Zero-shot stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5249–5260 (2025)
51. Wu, Z., Wang, Y., Wen, Y., Zhang, Z., Wu, B., Tang, H.: Stereoadapter: Adapting stereo depth estimation to underwater scenes. arXiv preprint arXiv:2509.16415 (2025)
52. Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21919–21928 (2023)
53. Xu, G., Wang, X., Zhang, Z., Cheng, J., Liao, C., Yang, X.: Igev++: Iterative multi-range geometry encoding volumes for stereo matching. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
54. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv preprint arXiv:2406.09414 (2024)
55. Yuh, J., West, M.: Underwater robotics. Advanced robotics **15**(5), 609–639 (2001)
56. Žbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. Journal of Machine Learning Research **17**(65), 1–32 (2016)
57. Zhang, F., You, S., Li, Y., Fu, Y.: Atlantis: Enabling underwater depth estimation with stable diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11852–11861 (2024)
58. Zhang, F., You, S., Li, Y., Fu, Y.: Learning rain location prior for nighttime deraining and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(10), 9169–9186 (2025)
59. Zhang, F., You, S., Li, Y., Fu, Y.: Atlantis++: Enabling underwater depth estimation with stable diffusion and beyond. International Journal of Computer Vision **134**(6), 260 (2026)
60. Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.: Ga-net: Guided aggregation net for end-to-end stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 185–194 (2019)
61. Zhang, L., Bao, C., Ma, K.: Self-distillation: Towards efficient and compact neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(8), 4388–4403 (2021)
62. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
63. Zhu, D., Huang, X., Huang, H., Zhou, H., Shao, Z.: Change3d: Revisiting change detection and captioning from a video modeling perspective. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24011–24022 (2025)
64. Zou, Y., Fu, Y., Zhang, Y., Zhang, T., Yan, C., Timofte, R.: Calibration-free raw image denoising via fine-grained noise estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
65. Zwilgmeyer, P.G.O., Yip, M., Teigen, A.L., Mester, R., Stahl, A.: The varos synthetic underwater data set: Towards realistic multi-sensor underwater data with ground truth. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3722–3730 (2021)