OPTI3R: GEOMETRY-LOCKED OPTICAL PAIRING AND RELIABILITY-WEIGHTED CROSS-VIEW SUPERVISION FOR UNDERWATER 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Attenuation and backscatter change underwater image appearance even when the scene geometry stays fixed. These changes can lead feed-forward models to predict inconsistent depth and shape. To reduce this effect, we propose Opti3R, which trains on Geometry-Locked Optical Pairing (GLOP): paired reference and optically perturbed images with consistent geometry targets. We further use teacher correspondences to supervise cross-view consistency with continuous visibility-confidence-reprojection (VCR) edge weighting, referred to as VCR Graph supervision (VCRG). Our proposed Axial-Channel Geometry Adapter (ACGA) adds identity-initialized multiscale residual blocks to the dense depth and point heads. These blocks take dense features as input and are optimized jointly with the GLOP and VCRG objectives. Raw inference achieves absolute relative depth errors of 0.0838 on SQUID and 0.0496 on FLSea-VI. GLOP gives lower reconstruction error than generic and unpaired augmentation. Opti3R also shows slower growth in geometry drift under held-out optical conditions, and higher VCR edge weights are associated with lower independent geometric disagreement. These results support geometry-locked optical pairing and continuously weighted cross-view supervision for underwater reconstruction.

1 INTRODUCTION

Underwater images undergo wavelength-dependent attenuation and scattering that produce color casts, contrast loss, and blurred details (Pucci & Martinel, 2025; Zeng et al., 2026). These degradations vary with illumination, scene range, and suspended particles, so the same structure can exhibit markedly different appearances across views (Varghese & Rajagopalan, 2025; Liu et al., 2025). Recent enhancement systems address these effects through latent-space alignment, region-aware masks, and uncertainty-aware reconstruction (Liu et al., 2025; Wang et al., 2025; Zeng et al., 2026). Yet perceptual enhancement alone does not guarantee geometrically stable correspondences or reconstruction (Yang et al., 2025a; Huang et al., 2025; Álvarez Tuñón & Graßhof, 2026). An underwater reconstruction model must therefore separate optical variation from changes in scene geometry (Ren et al., 2026; Xu et al., 2026a; Liu et al., 2026b).

Feed-forward geometry models predict depth, cameras, and point maps from one or more images without scene-specific optimization (Yang et al., 2025c; Keetha et al., 2026). Recent underwater variants use synthetic degradation, optical conditioning, or cross-view supervision to adapt these models to underwater imagery (Ren et al., 2026; Xu et al., 2026a; Liu et al., 2026b; Shimano & Fuse, 2026). Such adaptation must respond to degraded observations while preserving the geometric relationships learned before adaptation (Ren et al., 2026; Liu et al., 2026b).

Two problems remain. Appearance augmentation alone does not require the model to predict the same geometry before and after an optical change, leaving it free to interpret appearance variation as changes in depth or shape (Ren et al., 2026; Xu et al., 2026a). Cross-view correspondences also have different reliability (Ren et al., 2026). Binary masks are commonly constructed by thresholding or clustering scene evidence into foreground/background or valid/invalid regions; for example, depth-based geometric filtering retains pixels that satisfy a foreground or multi-view consistency test (Wang et al., 2025; Varghese & Rajagopalan, 2025; Ren et al., 2026). For correspondence learning, however,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

a hard mask gives every retained match the same weight and discards matches just below its threshold, including uncertain but informative evidence in low-contrast regions (Ren et al., 2026).

To address these problems, we propose Opti3R. Geometry-Locked Optical Pairing (GLOP) creates paired reference and optically perturbed images of the same scene, keeping the pixel coordinates and cameras unchanged. The teacher predicts geometry from the reference image, and the student learns to predict consistent geometry from the perturbed image. This gives the model a direct training constraint on how its predictions should behave under optical changes.

We further introduce VCRG supervision using teacher correspondences. Visibility checks remove invalid matches, while endpoint confidence and reprojection consistency determine continuous VCR edge weights for the remaining correspondences. To adapt dense features under these losses, we insert ACGA blocks into the depth and point heads. Each block computes multiscale spatial and channel corrections and starts as an identity mapping. Our contributions are:

- We propose Geometry-Locked Optical Pairing (GLOP), which enforces consistent geometry predictions for reference and optically perturbed images of the same scene.
- We propose Visibility-Confidence-Reprojection Graph supervision (VCRG), which assigns continuous three-factor weights to teacher correspondences.
- We propose the Axial-Channel Geometry Adapter (ACGA), an identity-initialized multiscale residual block optimized under the GLOP and VCRG objectives.

We evaluate depth and reconstruction quality, compare GLOP with generic augmentation, and calibrate VCR edge weights against independent geometric disagreement. We also compare ACGA with a residual adapter of similar size. Matched inference measurements and a five-run training-budget study report computational cost and training stability.

2 Related Work

Feed-forward geometric reconstruction. DUSt3R predicts pairwise point maps in a common coordinate frame, MASt3R augments this representation with local matching features, and Fast3R extends joint prediction to large image collections (Wang et al., 2024; Leroy et al., 2024; Yang et al., 2025b). VGGT jointly predicts cameras, depth, point maps, and tracks, whereas MapAnything incorporates optional geometric inputs to recover metric structure (Yang et al., 2025c; Keetha et al., 2026). Underwater extensions adapt this feed-forward paradigm in complementary ways: Enhanced DUSt3R uses shallow-water simulation, Wat3R learns from unlabeled video through teacher–student consistency, WAT3R introduces geometry-constrained degradation adaptation, and Water-VGGT conditions geometric reasoning on water optics (Shimano & Fuse, 2026; Ren et al., 2026; Xu et al., 2026a; Liu et al., 2026b). AquaFlow further combines foundation-model point maps with medium-aware Gaussian initialization for streaming reconstruction (Xu et al., 2026b). Wat3R uses teacher–student cross-view supervision with discrete region filtering, whereas Opti3R introduces geometry-locked optical pairs and assigns continuous correspondence weights jointly from visibility, endpoint confidence, and reprojection consistency. Opti3R retains the feed-forward setting while explicitly coupling paired optical invariance with graded cross-view evidence.

Underwater geometry estimation. Physics-guided underwater MVS addresses the air-to-water domain gap through synthetic image formation and end-to-end dense reconstruction (Yang et al., 2025a). A parallel line embeds the water medium into neural rendering: WaterSplatting separates explicit Gaussian geometry from a volumetric scattering field, while DualPhys-GS, UW-3DGS, and WaterClear-GS introduce attenuation–scattering pathways, physics-aware pruning, or optical parameters within Gaussian primitives (Li et al., 2025a;b; Xing et al., 2025; Zhang et al., 2026). Restoration-driven R-Splatting and spatiotemporal MarineSTD-GS instead target cross-view illumination variation, caustics, and transient degradation, and Swimm3R couples medium-aware SfM with underwater splatting (Huang et al., 2025; Liu et al., 2026a; Kweon & Sattar, 2026). A controlled cross-regime study shows that water clarity and evaluation conditions can dominate the apparent ranking of underwater Gaussian-splatting systems (Álvarez Tuñón & Graßhof, 2026). These approaches motivate explicit optical modeling, but Opti3R focuses on preserving geometry under paired optical changes and weighting teacher correspondences continuously during feed-forward adaptation.

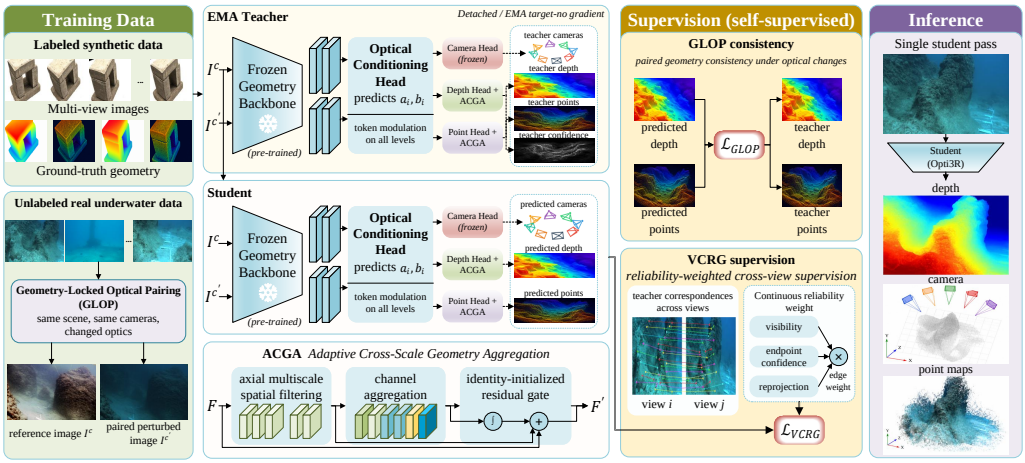


Figure 1: Opti3R training pipeline and ACGA placement.

3 METHOD

3.1 OVERVIEW

Task and overview. Given images $\mathcal{I} = \{I_i\}_{i=1}^N$, Opti3R predicts depth maps D_i, cameras G_i, and point maps P_i. We train with synthetic geometry labels and teacher targets from unlabeled underwater sequences. Figure 1 summarizes the training graph. The EMA teacher processes the reference image and supplies detached targets, whereas the student receives a geometry-preserving optical perturbation. GLOP consistency acts on depth and point outputs, while VCRG reprojection follows teacher correspondences whose edge weights indicate continuous reliability. ACGA is inserted before final upsampling in the dense heads. Correspondence search uses detached teacher cameras, and graph residuals use detached student cameras, so gradients pass through point prediction and upstream optical modulation but not through the projection cameras.

Shared optical conditioning and cameras. Let T_i^ℓ be tokens from the frozen geometry backbone at level ℓ. The trainable optical head takes the last token level and predicts scale and bias vectors a_i, b_i. We apply them to every level before the prediction heads:

$$\widehat{T}_i^\ell = T_i^\ell \odot \big(1 + 0.02\tanh(a_i)\big) + 0.02\tanh(b_i). \tag{1}$$

Here $\odot$ denotes elementwise multiplication. We broadcast the modulation over tokens and feed $\{\widehat{T}_i^\ell\}$ to the camera, depth, and point heads. Training the optical head can therefore change camera predictions even though the backbone and camera head are frozen. ACGA is placed later in the depth and point heads, so its outputs do not feed into the camera head. The supervised and teacher camera losses update optical modulation through the frozen camera head.

3.2 GEOMETRY-LOCKED OPTICAL PAIRING (GLOP)

Optical degradation can change depth predictions even when the scene stays fixed. GLOP reduces this effect by creating image pairs with different water conditions and the same reference geometry. The training loss then requires consistent predictions for the two images.

We generate paired observations with the standard underwater image-formation model. A reference observation I^c is associated with scene radiance J, depth D, and optical condition $c = (\beta^D, \beta^B, B^\infty)$, where the three-channel coefficients β^D and β^B control direct attenuation and backscatter, respectively, and B^∞ is the asymptotic background light. We vary these quantities while keeping scene radiance, depth, pixel coordinates, and cameras fixed, thereby changing appearance without changing the reference geometry. The symbol c' denotes the perturbed optical condition.

For real images, we apply an additional optical perturbation directly to the reference image. This training perturbation keeps pixel correspondences fixed but does not fully re-render the scene. We derive an optical path variable $\ell(x)$ from detached teacher depth $d(x)$ and use channel-dependent incremental coefficients to construct

$$\begin{aligned} I_k^{c'}(x) &= I_k^c(x) t_k^D(x) + A_k\big(1 - t_k^B(x)\big), \\ t_k^D(x) &= \exp(-\beta_k^{D,\Delta}\ell(x)), \qquad t_k^B(x) = \exp(-\beta_k^{B,\Delta}\ell(x)), \quad k \in \{r,g,b\}. \end{aligned} \tag{2}$$

Here A_k is the sampled background-light component. The dimensionless optical path and sampling distribution used in the controlled optical analysis are specified in Supplementary Section "Optical Sampling for the Controlled Analysis". The perturbation changes attenuation, backscatter, and background light but keeps the pixel grid and cameras fixed. The consistency loss can therefore compare predictions at the same pixel. Teacher depth errors in $d(x)$ can change the local perturbation strength. We detach this depth so that the model cannot reduce the consistency loss by changing the rendered input. With tied direct and backscatter coefficients, Equation 2 reduces to $I_k^{c'} = I_k^c t_k + A_k(1 - t_k)$.

Let f_θ^{geo} denote the student's dense depth and point outputs and $f_{\bar\theta}^{\mathrm{geo}}$ the corresponding targets from an EMA teacher (Tarvainen & Valpola, 2017). The student receives $I^{c'}$, whereas the teacher receives I^c, and all teacher outputs are detached during the student update. We initialize $\bar\theta = \theta$ and, after each optimizer step, update the teacher by the standard EMA rule $\bar\theta \leftarrow \mu\bar\theta + (1-\mu)\theta$ with $\mu = 0.996$.

Binary masks commonly turn scene or geometric evidence into discrete region decisions, such as foreground–background separation for underwater enhancement or exclusion of dynamic transients from geometric supervision (Wang et al., 2025; Varghese & Rajagopalan, 2025). In our paired setting, $m(x) = 1$ when the reference geometry is finite with positive depth and the perturbed RGB value is finite and lies in $[0,1]$; otherwise $m(x) = 0$. Synthetic pairs also use the reference geometry-validity mask. The operator $\mathcal{A}$ expresses corresponding dense outputs in a common coordinate and scale convention. We detach both the mask and rendered student input. The GLOP consistency loss is

$$\mathcal{L}_{\mathrm{GLOP}} = \frac{\sum_x m(x)\rho_\delta\Big(\mathcal{A}f_\theta^{\mathrm{geo}}(I^{c'})(x) - \mathrm{sg}\big[\mathcal{A}f_{\bar\theta}^{\mathrm{geo}}(I^c)(x)\big]\Big)}{\sum_x m(x) + \epsilon}. \tag{3}$$

Here sg stops gradients and $\epsilon > 0$ stabilizes normalization. The componentwise Huber penalty ρ_δ is quadratic for residual magnitudes up to δ and linear beyond that point. We set $\delta = 1.0$ for normalized geometric residuals and sum over vector components. For each image, depth consistency aligns the student depth D_s to the teacher depth D_t and normalizes it by a detached robust teacher scale:

$$\begin{aligned} (a_D, b_D) &= \arg\min_{a,b} \sum_x m(x)\big(aD_s(x) + b - D_t(x)\big)^2, \\ s_D &= \mathrm{sg}\big[\mathrm{median}_{x:m(x)=1}\, D_t(x)\big], \\ \widetilde{D}_s &= \frac{a_D D_s + b_D}{s_D}, \qquad \widetilde{D}_t = \frac{D_t}{s_D}. \end{aligned} \tag{4}$$

The fitted scale, shift, and normalization remain fixed during backpropagation. For point consistency, $\mathcal{A}$ first expresses the student and teacher point maps in a common camera frame. Each map is then divided by the detached median of its finite positive point norms over the valid support, using one scalar for all three coordinates. This makes depth and point residuals dimensionless; graph residuals remain in processed-image pixels. Empty valid support contributes zero loss, and maps without a finite positive scale are excluded. Supplementary Section "Implementation Details" summarizes the training settings.

This loss reduces prediction drift under optical changes. The shuffled-pair ablation in Section 4.5 breaks scene correspondence while keeping the marginal GLOP distribution unchanged.

3.3 Visibility-Confidence-Reprojection Graph Supervision (VCRG)

Different underwater views may retain useful texture on different parts of the same surface. Binary confidence masks can discard moderately reliable correspondences that still help constrain geometry. VCRG first removes geometrically invalid correspondences, then assigns continuous visibility-confidence-reprojection weights to valid ones.

We represent views as graph nodes and directed correspondences as edges (i, j) with $i \neq j$. For branch $b \in \{\theta, \bar{\theta}\}$, $G_i^b = (K_i^b, R_i^b, t_i^b)$ contains the intrinsics and the transform from the branch's common point frame to camera i, and $\pi(P; G)$ denotes transformation followed by perspective projection. For each source pixel $x \in \Omega_i$, we project the detached teacher point $P_i^{\bar{\theta}}(x)$ into view j with $G_j^{\bar{\theta}}$ to obtain $u_{ij}^t(x)$. The corresponding student point $P_i^{\theta}(x)$ is projected with detached camera G_j^{θ} to obtain $u_{ij}^s(x)$. Correspondence construction is therefore fixed during backpropagation, while the reprojection residual updates the student point prediction.

Around the rounded teacher projection, $\mathcal{N}(u_{ij}^t(x))$ contains valid integer pixels in Ω_j within Chebyshev distance two, yielding a 5×5 candidate window. Among candidates with finite teacher geometry and positive camera depth, $\widehat{x}_{ij}(x)$ is the target pixel whose teacher 3D point has the smallest Euclidean distance to $P_i^{\bar{\theta}}(x)$. An empty candidate set produces no valid edge. The student reprojection residual is $r_{ij}(x) = \|u_{ij}^s(x) - \widehat{x}_{ij}(x)\|_2$, measured in processed-image pixels.

The dense prediction heads output a confidence logit alongside geometry. Let $s_i^{\bar{\theta}}(x)$ be the EMA-teacher logit and define the bounded endpoint reliability as $q_i^{\theta}(x) = \sigma(s_i^{\bar{\theta}}(x))$, where σ is the sigmoid. The native dense-prediction objective trains confidence, and the EMA update transfers it to the teacher. Non-finite confidence values invalidate an edge. For a matched edge, the two endpoint scores are combined by their geometric mean, $q_{ij}(x) = \sqrt{q_i^{\bar{\theta}}(x) q_j^{\bar{\theta}}(\widehat{x}_{ij}(x))}$.

The teacher visibility test rejects projections outside either image and points behind a camera. The detached indicator $v_{ij}(x)$ is one only when the forward teacher projection, the return teacher projection, and the student projection are finite, in-image, and have positive camera depth, with the return projection lying within two processed-image pixels of the source pixel. We then define the return check and continuous edge weight as

$$\begin{aligned} \left\| \mathrm{sg}\left[\pi\left(P_j^{\bar{\theta}}(\widehat{x}_{ij}(x)); G_i^{\bar{\theta}} \right) \right] - x \right\|_2 &\leq 2 \text{ pixels}, \\ w_{ij}(x) &= \mathrm{sg}\left[v_{ij}(x) q_{ij}(x) \exp\left(-\frac{r_{ij}(x)}{\tau} \right) \right]. \end{aligned} \tag{5}$$

We set $\tau = 1.0$ pixel, in the same units as r_{ij}. The graph loss is the normalized weighted reprojection penalty,

$$\mathcal{L}_{\mathrm{VCRG}} = \frac{\sum_{i \neq j} \sum_x w_{ij}(x) \rho_\delta(r_{ij}(x))}{\sum_{i \neq j} \sum_x w_{ij}(x) + \epsilon}, \tag{6}$$

where the sums run over source-view pixels $x \in \Omega_i$, and ρ_δ is the componentwise Huber penalty with $\delta = 1.0$ pixel. Invalid correspondences have zero weight and are excluded from both sums. The detached weight prevents the graph objective from updating confidence through the weighting function. Gradients through r_{ij} update student point prediction; detaching G_j^{θ} stops gradients through the projection camera, while the optical head still receives graph gradients through point predictions. The hard-mask ablation in Section 4.5 replaces VCR edge weights with binary selection.

3.4 Axial-Channel Geometry Adapter (ACGA)

Underwater degradation affects both local edges and broad low-contrast regions, and it can change which feature channels remain informative. We therefore insert ACGA blocks into the dense depth and point heads. Each block receives a fused feature map and produces a residual correction through multiscale axial filtering and channel aggregation. The VCR edge weights modulate the graph-loss gradients reaching the point-head ACGA block, while GLOP consistency trains both dense-head blocks. ACGA itself consumes only the feature map and requires no explicit reliability input.

Let $z \in \mathbb{R}^{B \times C \times H \times W}$ denote a fused feature before final upsampling in a depth or point head. The ACGA output is

$$z' = z + \tanh(\alpha)\big(\mathcal{P}_\phi(z) - z\big), \qquad \alpha_0 = 0. \tag{7}$$

The transformation $\mathcal{P}_\phi$ combines axial multiscale filtering with channel-wise query/key/value aggregation. We train the blocks in the geometric heads using Equations 3 and 6, as described above.

The transformation $\mathcal{P}_\phi$ uses shared multiscale axial filters followed by pooled channel-wise query/key/value aggregation. Its exact tensor operations, dimensions, and normalization are given in

Supplementary Section "Implementation Details"; the construction preserves the $B \times C \times H \times W$ feature shape.

At initialization, $\alpha = 0$, so $z' = z$ and ACGA leaves the pretrained head output unchanged. Initially, $\partial z'/\partial \phi = 0$, while $\partial z'/\partial \alpha = \mathcal{P}_\phi(z) - z$, allowing the scalar gate to change before the feature-transformation parameters. The depth and point heads have separate ACGA parameters and scalar gates; both use $C = 128$ and eight heads before final upsampling. GLOP consistency trains both blocks, whereas VCRG reprojection additionally trains the point-head block. We evaluate ACGA without these consistency objectives and measure its effect by comparing GLOP plus VCRG (GLOP + VCRG) with the same objectives plus ACGA (GLOP + VCRG + ACGA).

3.5 Joint Learning Objective

We train one predictor with synthetic geometry labels and EMA-teacher targets, together with the GLOP and VCRG consistency losses. The complete objective also includes auxiliary supervision for optical parameters:

$$\mathcal{L} = \mathcal{L}_{\text{supervised}} + \lambda_u \mathcal{L}_{\text{teacher}} + \lambda_i \mathcal{L}_{\text{GLOP}} + \lambda_g \mathcal{L}_{\text{VCRG}} + \lambda_o \mathcal{L}_{\text{optics}}. \tag{8}$$

Synthetic labeled observations use $\mathcal{L}_{\text{supervised}} = 5\mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{point}}$. The depth and point objectives include regression, confidence-weighted regression with logarithmic confidence regularization, and spatial regularization. Thus, confidence is learned through geometric prediction error.

The coefficient λ_u weights $\mathcal{L}_{\text{teacher}}$, which supervises camera, depth, and point predictions on unlabeled real images using detached EMA targets. The coefficient λ_i weights $\mathcal{L}_{\text{GLOP}}$ on paired reference and perturbed images (Equation 3). The coefficient λ_g weights the cross-view reprojection loss $\mathcal{L}_{\text{VCRG}}$ (Equation 6).

The coefficient λ_o weights optical-parameter regression. For samples $n \in \mathcal{S}_o$ with known optical target vectors c_n^* and predictions $\widehat{c}_n$, $\mathcal{L}_{\text{optics}}$ is the mean componentwise Huber error between $\widehat{c}_{n,k}$ and $c_{n,k}^*$ over optical components k, with ϵ added to the normalizing count. It is zero when a batch contains no optical targets. We set $\lambda_u = 1.0$, $\lambda_i = 1.0$, $\lambda_g = 0.5$, and $\lambda_o = 0.1$. Supplementary Section "Computational Efficiency" reports trainable parameter counts.

GLOP consistency updates depth and point prediction, while VCRG reprojection updates point prediction with fixed camera tensors. These losses train the corresponding ACGA blocks, dense layers, and upstream optical modulation. The backbone weights remain fixed.

Raw inference requires one geometry prediction. Horizontal flipping preserves scene depth after mapping the output back to the original coordinates, but changes the left–right orientation of appearance and structure. We therefore use flip ensembling to test prediction stability under this symmetry, reduce orientation-specific artifacts, and quantify the resulting accuracy–cost trade-off under identical settings for both models. We predict depth from the original image and its horizontal flip, map the flipped prediction back, and fit a least-squares scale and shift between the two predictions on Ω_q. The ensemble combines 0.75 times the original prediction with 0.25 times the aligned flip-back prediction. The support Ω_q retains finite pairs within the first-to-ninety-ninth percentile ranges of both predictions. This prediction-to-prediction alignment is separate from alignment to benchmark reference depth.

4 Experiments

4.1 Experimental Setup

We evaluate depth, reconstruction, and pose, then study GLOP pairing, prediction drift under held-out optical conditions, VCR edge weighting, and ACGA. The ACGA comparison includes a residual adapter with a similar parameter count.

Benchmarks and metrics. We evaluate monocular depth on SQUID (Berman et al., 2021) and FLSea-VI (Randall & Treibitz, 2023). SQUID contains 57 stereo pairs, which we evaluate as 114 monocular images; for FLSea-VI, we use Coral Table Loop and U-Canyon. We evaluate multi-view

Table 1: Monocular depth on SQUID and FLSea-VI; lower Rel and higher δ_1 are better.

		SQUID		FLSea-VI	
Method	Inference	$\mathrm{Rel}\downarrow$	$\delta_1\uparrow$	$\mathrm{Rel}\downarrow$	$\delta_1\uparrow$
VGGT	Raw	0.1940	0.7030	0.1070	0.8880
Depth Anything V2	Raw	0.0990	0.9000	0.0690	0.9580
MapAnything	Raw	0.1040	0.8980	0.0860	0.9510
Wat3R	Raw	0.1070	0.8930	0.0610	0.9710
Opti3R	Raw	**0.0838**	**0.9250**	**0.0496**	**0.9790**
Wat3R	Flip	0.1048	0.8962	0.0605	0.9715
Opti3R	Flip	**0.0847**	**0.9262**	**0.0509**	**0.9794**

depth on Sea-thru (Akkaynak & Treibitz, 2019) and FLSea-Stereo (Randall & Treibitz, 2023) with ten input views. Absolute relative depth error Rel averages $|D - D^*|/D^*$ over valid pixels. Threshold accuracy δ_1 is the fraction of pixels satisfying $\max(D/D^*, D^*/D) < 1.25$. We resize inputs with a longest-side target of 518 pixels and evaluate depth after scale–shift alignment on valid reference pixels.

We evaluate Water3D (Ren et al., 2026) reconstruction on 42 scenes with 20 uniformly selected views per scene. Accuracy is the nearest-neighbor distance from reconstructed points to reference points, and completeness measures the reverse direction. Overall is their arithmetic mean, with lower values being better for all three metrics. We report depth-plus-camera and direct-point reconstruction separately. We evaluate relative pose on SeaThru-NeRF (Levy et al., 2023) using AUC at angular thresholds of 5°, 15°, and 30°, reported as fractions.

Training data and splits. We use labeled DTU for geometry supervision and a subset of the FLSea Canyons data, excluding U-Canyon, for unlabeled real-image training. U-Canyon is strictly excluded from training, and is used only for evaluation. DTU scans 37 and 40 are reserved for validation, and the remaining configured scans are used for training. Real sequences must contain at least 24 frames, and the configuration uses up to eight teacher views. We define scene and sequence splits before generating interventions so that related frames and optical variants stay in the same partition. We select the model using validation depth error on the held-out DTU scans. SQUID, FLSea-VI, Sea-thru, FLSea-Stereo, Water3D, and SeaThru-NeRF are used for the evaluations described above. Optical perturbations use DTU geometry for labeled samples and detached teacher depth for real images.

Training and comparisons. We train for 19,200 optimizer updates using AdamW with learning rate 10^{-4} and weight decay 0.05. We use linear warm-up for 5% of the updates (960 updates at the primary budget), followed by cosine decay. Each update uses eight sequences with eight views each, giving 64 effective views without gradient accumulation. We use bfloat16 precision and clip the gradient norm at 1.0. The geometry backbone and camera head are frozen. We train the ACGA blocks, dense depth and point heads, and optical head. Confidence is predicted by the dense heads. Camera predictions can change through the optical modulation in Equation 1. Supplementary Section "Implementation Details" lists the training settings.

The main benchmark and component-ablation runs use a fixed seed. The training-budget study reports means and sample standard deviations from five runs with seeds $\{0, 1, 2, 3, 4\}$. Component comparisons use the primary budget, and the budget study also includes 38,400 and 76,800 updates. Detailed training-budget results are reported in Supplementary Section "Training-Budget Analysis". For monocular depth, we compare primarily with VGGT (Yang et al., 2025c), Depth Anything V2 (Yang et al., 2024), MapAnything (Keetha et al., 2026), and Wat3R (Ren et al., 2026). Additional comparisons with underwater-specific depth methods and recent feed-forward geometry models are reported in Supplementary Section "Additional Comparisons". Raw inference is the primary setting, and matched horizontal-flip inference is reported separately for Wat3R and Opti3R. Supplementary Section "Evaluation Datasets" gives additional evaluation details.

4.2 Main Results

Table 2: Multi-view depth with ten input views and scale–shift alignment. Both methods use the same evaluation protocol within each dataset.

	Sea-thru		FLSea-Stereo	
Method	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑
Wat3R	0.0680	0.9820	0.0830	0.9430
Opti3R	**0.0592**	**0.9850**	**0.0695**	**0.9570**

Table 3: Water3D reconstruction and SeaThru-NeRF relative pose; lower reconstruction distance and higher pose AUC are better.

Task	Method	Accuracy↓	Completeness↓	Overall↓
Depth+Cam	Wat3R	0.4460	0.3660	0.4060
Depth+Cam	Opti3R	0.3650	0.2990	0.3320
Point	Wat3R	0.4440	0.4090	0.4265
Point	Opti3R	0.3800	0.3480	0.3640
		AUC@5↑	AUC@15↑	AUC@30↑
SeaThru-NeRF	Wat3R	0.5400	0.8200	0.9060
SeaThru-NeRF	Opti3R	**0.5870**	**0.8470**	**0.9200**

Monocular depth and matched inference. Opti3R had the lowest raw depth error among the methods in Table 1. Compared with Wat3R, raw Rel was lower by 21.68% on SQUID and 18.69% on FLSea-VI. On SQUID, the reduction relative to Depth Anything V2 was 15.35%.

With flip inference applied to both models, Opti3R reduced relative error by 19.18% on SQUID and 15.87% on FLSea-VI compared with Wat3R. The reductions were similar to those under raw inference.

Multi-view depth, reconstruction, and pose. Opti3R reduced multi-view depth error by 12.94% on Sea-thru and 16.27% on FLSea-Stereo (Table 2). It also improved depth accuracy on both datasets when multiple views of the same scene were available.

Opti3R had lower Water3D distances than Wat3R in the reported reconstruction evaluation (Table 3). The component comparisons use the same reconstruction metrics and are reported in Supplementary Section "Additional Ablation Tables".

Pose AUC was higher than Wat3R at all three thresholds. At 5°, it increased from 0.5400 to 0.5870, a gain of 4.70 percentage points. These results evaluate the full predictor, whose camera estimates depend on the optical conditioning in Equation 1.

Qualitative monocular-depth comparisons use the same crop, valid mask, depth range, and error-scale protocol within each row (Figure 2).

4.3 Geometry-Locked Pairing Analysis

GLOP and appearance augmentation. Under matched initialization, data, budget, validation selection, and raw inference, GLOP outperformed both generic photometric augmentation and unpaired optical perturbation on SQUID, FLSea-VI, and Water3D. The full control table and augmentation definitions are provided in Supplementary Section "Additional Ablation Tables".

Optical distribution shift. We train on normalized attenuation coordinates in $[0, 0.8)$ and evaluate on the held-out range $[0.8, 1]$. The renderer, severity partitions, drift definition, and numerical values are given in Supplementary Section "Optical Drift and Correspondence Reliability Measurements". Opti3R shows slower depth-drift growth than Wat3R and generic augmentation, including in the held-out range.

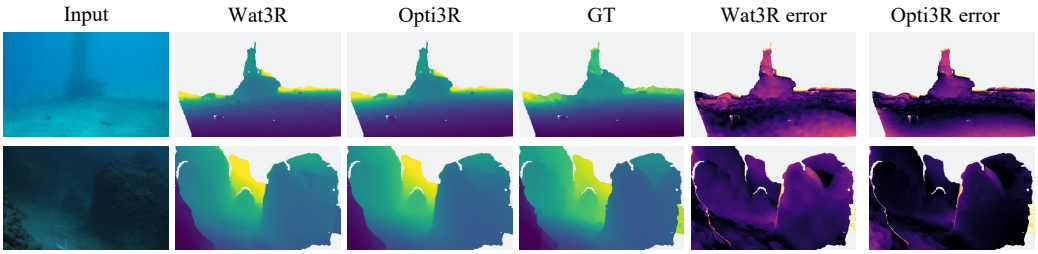


Figure 2: Qualitative monocular-depth comparisons on SQUID and FLSea-VI.

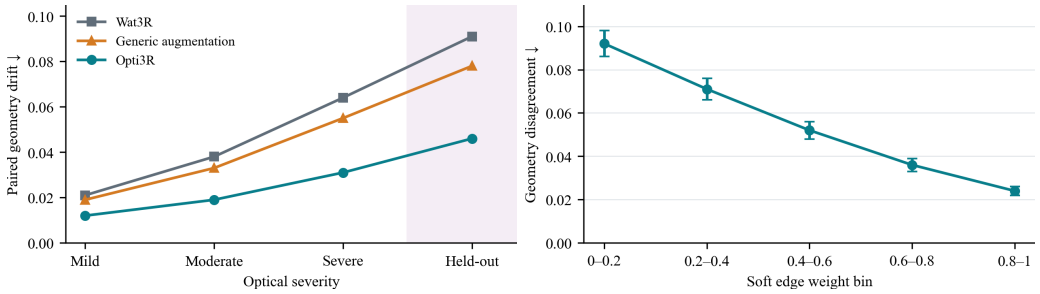


Figure 3: Mechanism analyses. Left: paired depth drift across optical severity levels, including the held-out range. Right: VCR edge weight versus independent point disagreement on the two held-out DTU scenes.

4.4 Calibration of VCR Edge Weights

VCRG assigns continuous weights to valid correspondences; replacing them with a hard mask increased Water3D distance from 0.3320 to 0.3650 (Supplementary Section "Additional Ablation Tables"). On two held-out DTU scenes, disagreement decreased monotonically across weight bins, with mean edge-level rank correlation $\rho = -0.71$. The calibration protocol, equation, and uncertainty calculation are given in Supplementary Section "Optical Drift and Correspondence Reliability Measurements"; the two mechanism plots are shown in Figure 3.

4.5 Ablation and ACGA Control

We compare Opti3R variants under the same adaptation settings, using 19,200 optimizer updates, a fixed seed, and raw inference. Geometry-Locked Optical Pairing (GLOP) and VCRG are combined in GLOP + VCRG and GLOP + VCRG + ACGA, which differ only in the use of ACGA. The plain-residual control uses the same insertion points and identity-initialized gates.

The hard-mask and shuffled-pair definitions, together with the detailed component comparisons, are given in Supplementary Section "Additional Ablation Tables". In brief, same-scene GLOP, continuous VCR weighting, and ACGA each improve the matched evaluation, while the five-seed training-budget study supports 19,200 updates (Supplementary Section "Training-Budget Analysis"). Computational measurements are reported in Supplementary Section "Computational Efficiency"; raw Opti3R adds 8.54% latency over Wat3R with matched heads.

5 Discussion and Conclusion

We presented Opti3R for feed-forward underwater 3D reconstruction. GLOP separates optical variation from scene geometry, VCRG retains graded cross-view evidence, and ACGA adds lightweight dense-head adaptation capacity. Experiments show gains from paired optical consistency, continuous correspondence weighting, and ACGA under joint optimization.

Reproducibility Statement

We provide the information needed to reproduce our method and experiments throughout the paper and supplementary materials. Sections 3.2–3.5 specify the proposed GLOP framework, including the teacher update, correspondence matching, VCR edge weights, ACGA, and the complete training objective. Section 4 describes the datasets, evaluation protocols, metrics, and baseline comparisons. The supplementary materials provide implementation and optimization details, trainable modules, dataset evaluation units, repeated-run aggregation procedures, and computational measurements. We additionally provide anonymous source code and supporting scripts at `https://anonymous.4open.science/r/opti3r-8FFB/` to facilitate reproduction of the reported experiments.

AI Use Statement

Generative AI tools were used only to aid and polish the writing.

References

Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 1682–1691. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00178. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Akkaynak_Sea-Thru_A_Method_for_Removing_Water_From_Underwater_Images_CVPR_2019_paper.html.

Dana Berman, Deborah Levy, Shai Avidan, and Tali Treibitz. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(8):2822–2837, 2021. doi: 10.1109/TPAMI.2020.2977624. URL https://doi.org/10.1109/TPAMI.2020.2977624.

Guoxi Huang, Haoran Wang, Zipeng Qi, Wenjun Lu, David Bull, and Nantheera Anantrasirichai. From restoration to reconstruction: Rethinking 3d gaussian splatting for underwater scenes, 2025. URL https://doi.org/10.48550/arXiv.2509.17789.

Nikhil Varma Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian A. Scherer, and Peter Kontschieder. Mapanything: Universal feed-forward metric 3d reconstruction; map-anything.github.io. In *International Conference on 3D Vision, 3DV 2026, Vancouver, BC, Canada, March 20-23, 2026*, pp. 499–509. IEEE, 2026. doi: 10.1109/3DV69130.2026.00054. URL https://doi.org/10.1109/3DV69130.2026.00054.

Minseong Kweon and Junaed Sattar. Swimm3r: Splatting with medium-aware sfm for underwater 3d reconstruction, 2026. URL https://doi.org/10.48550/arXiv.2608.00950.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXII*, volume 15130 of *Lecture Notes in Computer Science*, pp. 71–91. Springer, 2024. doi: 10.1007/978-3-031-73220-1_5. URL https://doi.org/10.1007/978-3-031-73220-1_5.

Deborah Levy, Amit Peleg, Naama Pearl, Dan Rosenbaum, Derya Akkaynak, Simon Korman, and Tali Treibitz. Seathru-nerf: Neural radiance fields in scattering media. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 56–65. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00014. URL https://doi.org/10.1109/CVPR52729.2023.00014.

Huapeng Li, Wenxuan Song, Tianao Xu, Alexandre Elsig, and Jonas Kulhanek. Watersplatting: Fast underwater 3d scene reconstruction using gaussian splatting. In *International Conference on 3D Vision, 3DV 2025, Singapore, March 25-28, 2025*, pp. 969–978. IEEE, 2025a. doi: 10.1109/3DV66043.2025.00094. URL https://doi.org/10.1109/3DV66043.2025.00094.

Jiachen Li, Guangzhi Han, Jin Wan, Yuan Gao, and Delong Han. Dualphys-gs: Dual physically-guided 3d gaussian splatting for underwater scene reconstruction. *Comput. Graph.*, 133:104405, 2025b. doi: 10.1016/J.CAG.2025.104405. URL https://doi.org/10.1016/j.cag.2025.104405.

Shaohua Liu, Ning Gao, Zuoya Gu, Hongkun Dou, Yue Deng, and Hongjue Li. Spatiotemporal degradation-aware 3d gaussian splatting for realistic underwater scene reconstruction, 2026a. URL https://doi.org/10.48550/arXiv.2604.23551.

Yifan Liu, Chuanbo Zhu, Ke Luo, and Jincai Chen. Underwater vision clearing with latent space alignment for marine scene enhancement. *Displays*, 90:103094, 2025. doi: 10.1016/J.DISPLA.2025.103094. URL https://doi.org/10.1016/j.displa.2025.103094.

Yifan Liu, Bin Peng, Mingke Xing, Tong Hua, and Jincai Chen. Water-vggt: Leveraging visual geometry and water-optics for robust 3d reconstruction in underwater environments. *Ocean Engineering*, 364:127044, 2026b. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2026.127044. URL https://www.sciencedirect.com/science/article/pii/S0029801826028787.

Rita Pucci and Niki Martinel. CE-VAE: capsule enhanced variational autoencoder for underwater image enhancement. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025*, pp. 2113–2123. IEEE, 2025. doi: 10.1109/WACV61041.2025.00212. URL https://doi.org/10.1109/WACV61041.2025.00212.

Yelena Randall and Tali Treibitz. Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets. *CoRR*, abs/2302.12772, 2023. doi: 10.48550/ARXIV.2302.12772. URL https://doi.org/10.48550/arXiv.2302.12772.

Jiangwei Ren, Xingyu Jiang, Zijie Song, Wei Xu, Hongkai Lin, Dingkang Liang, and Xiang Bai. Wat3r: Underwater 3d geometry learning without annotations, 2026. URL https://arxiv.org/abs/2607.08772.

T. Shimano and T. Fuse. Enhanced dust3r for underwater 3d reconstruction in shallow water environments. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIX-B2-2026:997–1004, 2026. doi: 10.5194/isprs-archives-XLIX-B2-2026-997-2026. URL https://isprs-archives.copernicus.org/articles/XLIX-B2-2026/997/2026/.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=ry8u21rtl.

Nisha Varghese and A. N. Rajagopalan. Sea-ing in low-light. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 16629–16640. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01550. URL https://openaccess.thecvf.com/content/CVPR2025/html/Varghese_Sea-ing_in_Low-light_CVPR_2025_paper.html.

Hao Wang, Kevin Köser, and Peng Ren. Large foundation model empowered discriminative underwater image enhancement. *IEEE Trans. Geosci. Remote. Sens.*, 63:1–17, 2025. doi: 10.1109/TGRS.2025.3525962. URL https://doi.org/10.1109/TGRS.2025.3525962.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. Dust3r: Geometric 3d vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 20697–20709. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01956. URL https://doi.org/10.1109/CVPR52733.2024.01956.

Wenpeng Xing, Jie Chen, Zaifeng Yang, Changting Lin, Jianfeng Dong, Chaochao Chen, Xun Zhou, and Meng Han. Uw-3dgs: Underwater 3d reconstruction with physics-aware gaussian splatting, 2025. URL https://arxiv.org/abs/2508.06169.

Jiayi Xu, Jiahao Lu, Ziqiang Zheng, Yihao Tan, Yaolong Zhu, Yuan Liu, and Sai-Kit Yeung. Wat3r: Feedforward underwater 3d reconstruction, 2026a. URL https://arxiv.org/abs/2607.21023.

Yingxiang Xu, Kerui Ren, Wenqi Guo, Changjian Jiang, Tao Lu, Linning Xu, and Mulin Yu. Aquaflow: A monocular gaussian splatting slam for underwater streaming reconstruction, 2026b. URL https://arxiv.org/abs/2608.22906.

Guidong Yang, Junjie Wen, Lei Lei, Benyun Zhao, Qingxiang Li, Xi Chen, Zhi Gao, and Ben M. Chen. Toward end-to-end underwater multi-view stereo for real-world dense scene reconstruction. *IEEE Trans. Geosci. Remote. Sens.*, 63:1–19, 2025a. doi: 10.1109/TGRS.2025.3630174. URL https://doi.org/10.1109/TGRS.2025.3630174.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21924–21935, 2025b.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 21924–21935. Computer Vision Foundation / IEEE, 2025c. doi: 10.1109/CVPR52734.2025.02042. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Fast3R_Towards_3D_Reconstruction_of_1000_Images_in_One_Forward_CVPR_2025_paper.html.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58-Abstract-Conference.html.

Xiuna Zeng, Jiaao Peng, Zhenqi Fu, Linyu Fan, Xiaotong Tu, Yue Huang, and Xinghao Ding. Accelerating adaptive diffusion and uncertainty modeling for underwater image enhancement. *IEEE Trans. Multim.*, 28:1065–1079, 2026. doi: 10.1109/TMM.2025.3632678. URL https://doi.org/10.1109/TMM.2025.3632678.

Xinrui Zhang, Yufeng Wang, Shuangkang Fang, Zesheng Wang, Dacheng Qi, and Wenrui Ding. Waterclear-gs: Optical-aware gaussian splatting for underwater reconstruction and restoration, 2026. URL https://arxiv.org/abs/2601.19753.

Olaya Álvarez Tuñón and Stella Graßhof. Gaussian splatting underwater: A controlled cross-regime study, 2026. URL https://arxiv.org/abs/2608.25483.