LinStereo: Linear-Complexity Global Attention for Multi-Scale Iterative Stereo Matching

Yiran Wang*, Oliver Turner, and Viorela Ila**

Australian Centre for Robotics (ACFR), School of Aerospace, Mechanical and Mechatronic Engineering (AMME), Faculty of Engineering, The University of Sydney, Sydney, NSW, Australia

Abstract. Existing Vision Foundation Model (VFM)-based iterative stereo pipelines under-exploit three information pathways: multi-scale backbone features are collapsed into single-level correlations, geometric priors remain untapped at initialization, and context propagates only locally. These gaps widen under degraded photometric cues, making underwater scenes a stringent generalization test. To address this, we propose LinStereo, built upon Depth Anything V3, whose core is a Position-Aware Linear Attention (PALA) module that replaces local recurrence with global aggregation at linear cost, propagating reliable estimates from well-matched regions into degraded areas while preserving disparity structure. PALA is made effective by two enabling components: Hierarchical Semantic Cost Volumes (HSCV), which supply scale-aligned correlations from the VFM feature hierarchy, and a Depth Prior Initialization (DPI) that converts monocular depth into a metrically calibrated warm start. LinStereo achieves state-of-the-art-level accuracy on standard benchmarks and strong cross-domain generalization, particularly on underwater scene where severe photometric degradation makes stereo matching particularly challenging, attaining the best overall accuracy with consistent gains (**28%** lower AbsRel on TartanAir-UW, **26%** on SQUID, a real-world underwater dataset). Code is available at `https://u7079256.github.io/LinStereo/`.

Keywords: Depth Estimation · Stereo Matching · Underwater Vision

1 Introduction

Stereo depth estimation recovers dense metric 3D geometry from calibrated image pairs and is fundamental for vision-based navigation, manipulation, and inspection, where both accuracy and inference speed directly affect performance [13]. Depth estimation remains difficult in textureless, occluded, or repetitive regions, which require strong contextual reasoning under limited computation. These challenges intensify underwater, where wavelength-dependent attenuation and volumetric scattering degrade visual signals, and limited annotated subsea data

* Corresponding author.
** Project lead.

causes models trained on synthetic or in-air datasets to suffer significant performance drops [1, 4, 27].

The core difficulty in dense stereo depth estimation lies in establishing reliable correspondences over large, ambiguous regions while keeping computation tractable. Some approaches construct explicit cost volumes and regularize them with 3D aggregation networks [7, 45]; others adopt iterative refinement, using a recurrent operator to progressively refine depth estimates by indexing a correlation volume [26, 41]. The iterative family has become dominant and has been further strengthened by replacing conventional encoders with vision foundation model (VFM) backbones to improve cross-domain robustness [20, 43] and by fusing monocular depth priors to regularize ill-posed regions [3, 53]. These efforts have substantially raised the accuracy ceiling, yet they focus on improving the inputs to the update loop; the update mechanism itself has received less attention. Crucially, VFM backbones provide semantically rich multi-scale features and implicit geometric priors for global reasoning, yet their effective use within the update loop remains underexplored.

We observe a growing mismatch between the capabilities of modern backbones and what the iterative update loop actually exploits. Even in recent VFM-based pipelines, three bottlenecks persist: updates reason primarily over local neighborhoods, multi-scale features are often reduced to single-level correlations, and the backbone's geometric priors remain largely unused at initialization. Although recent works address some of these aspects individually [3, 45, 53], as backbone capacity continues to grow, the backbone–update interface remains a key bottleneck.

LinStereo closes this gap with a redesigned decoder whose **core is Position-Aware Linear Attention (PALA)**: a global, linear-complexity update operator that, unlike local recurrent operators [26, 43], propagates information across the entire feature map at every refinement step. Two components make this global aggregation effective: **Hierarchical Semantic Cost Volumes (HSCV)** supply each refinement level with a correlation volume built from the VFM's matching-scale features, giving PALA semantically aligned matching evidence; and **Depth Prior Initialization (DPI)** converts the backbone's monocular depth into a disparity-space initialization, so PALA refines from a geometrically plausible state. LinStereo is thus a single PALA-centred decoder enabled by HSCV and DPI, rather than three independent modules.

These components address complementary aspects of the backbone–update interface and yield consistent improvements across all settings (Sec. 4). LinStereo achieves state-of-the-art accuracy among methods trained on comparable data on standard stereo benchmarks and demonstrates strong cross-domain generalization, particularly in underwater scenes with severe photometric degradation, achieving **28%** lower AbsRel on TartanAir-UW [40] and **26%** on SQUID [4].

The main contributions of this work are summarized as follows:

- **LinStereo**, an iterative stereo framework whose core is **PALA**, a position-aware linear-attention operator that replaces local recurrence with global aggregation at linear complexity. PALA is paired with **HSCV**, a hierarchical

cost-volume design that contributes both a strategy for using the backbone's multi-scale features and a new structure for each per-scale volume; and with **DPI**, which provides a warm start for refinement from monocular depth. Together, these reach state-of-the-art-level accuracy with significantly fewer refinement iterations at linear per-iteration cost.
– Trained only on SceneFlow with no underwater data, LinStereo achieves state-of-the-art performance on underwater stereo benchmarks without domain specific adaptation, attaining the best overall accuracy, demonstrating strong cross-domain generalization under severe photometric degradation.

The paper is complemented by open-source code and a new evaluation dataset with dense disparity ground truth and realistic subsea optical modeling, as described in the supplementary multimedia.

2 Related Work

Stereo Depth Estimation. Cost volume-based methods [7, 18, 21, 37, 47, 55] build explicit 3D/4D volumes regularized by 3D convolutions, but their cost scales with disparity range and single-scale construction limits multi-level semantic integration. RAFT-Stereo [26] introduced an iterative alternative: a recurrent operator repeatedly queries a fixed correlation volume to refine disparity through local updates. Subsequent works enrich this paradigm with cascaded correlation [23], geometry-encoded volumes [45], and frequency-adaptive units [41], yet the update operator remains spatially local, typically requiring tens of iterations for information to propagate across the image.
This locality becomes increasingly significant as vision foundation models (VFMs) enter the pipeline. FoundationStereo [43], DEFOM-Stereo [20], MGStereo [53], and Stereo Anywhere [3] leverage pretrained encoders [24, 30, 51, 52, 54] to extract globally rich features that substantially improve zero-shot generalization, yet these features are still consumed through a local recurrent loop, which creates an information utilization gap between backbone capacity and update reach. Our work targets this mismatch, enabling the iterative pipeline to better exploit backbone representations and converge in fewer iterations.

Efficient Stereo Depth Estimation. Reducing inference cost while maintaining accuracy is a central challenge in stereo depth estimation. Prior work primarily simplifies the architecture itself: pruning the disparity search space [12,22,36,44], replacing 3D cost aggregation with lightweight 2D alternatives [2,17,42,47,48,50], or eliminating explicit volumes entirely [39]. While achieving impressive speed, such simplifications can limit accuracy in challenging regions. Our method pursues efficiency from a complementary direction: rather than simplifying the architecture, we reduce the number of refinement iterations required for convergence, preserving rich contextual modeling while lowering inference time.

Underwater Depth Estimation. Underwater environments amplify the challenges of stereo depth estimation: wavelength-dependent attenuation, backscattering,

and refractive distortions severely degrade photometric consistency, texture contrast, and color fidelity across stereo views [1, 4, 10, 11]. These degradations directly weaken the correspondence cues that stereo methods rely on, making underwater scenes a particularly demanding setting for evaluating contextual reasoning capability. Domain-specific methods such as UWNet [56] introduce attention and cross-search modules tailored to underwater imagery but remain tightly coupled to specific degradation patterns. On the data side, annotated underwater stereo corpora are scarce: UWStereo [27] provides dense synthetic annotations but exhibits a sim-to-real gap, and existing datasets generally lack physically grounded underwater optical modeling paired with dense stereo ground truth. To address this data bottleneck, we introduce SeaStereo-Dataset, a physically rendered underwater stereo corpus with dense disparity annotations under realistic subsea optical conditions.

In summary, iterative stereo methods suffer from an information gap between powerful backbones and their local update interface, efficient methods sacrifice contextual capacity for speed, and underwater stereo remains limited by data scarcity. LinStereo addresses all three: it widens the update interface to leverage modern backbones, achieves efficiency through rapid convergence rather than architectural simplification, and contributes a large-scale underwater training and evaluation corpus.

3 The Proposed Method

3.1 Overview

We propose LinStereo, a stereo depth estimation framework that addresses the growing mismatch between VFM backbone capacity and iterative update utilization in stereo pipelines. Recent methods [20, 43] have demonstrated that VFM backbones significantly improve stereo depth estimation through richer feature representations, yet they inherit the narrow-bandwidth update architecture from pre-VFM designs [26, 45], creating a fundamental mismatch between a high-capacity encoder and a low-capacity iterative decoder. As illustrated in Fig. 1, LinStereo closes this gap with three components that each target a distinct information pathway: **Hierarchical Semantic Cost Volumes (HSCV)** exploit the VFM's multi-scale hierarchy to provide semantically matched correlation at every refinement level (*what* to match); **Depth Prior Initialization** converts the backbone's monocular depth into a geometrically informed warm start (*where* to start); and **Position-Aware Linear Attention (PALA)** replaces local convolutions with global context aggregation at linear cost (*how* to refine). These components target complementary information pathways and, as shown in Sec. 4, yield consistent gains across all evaluated settings. The following subsections present them in pipeline order: from feature construction through initialization to iterative refinement.

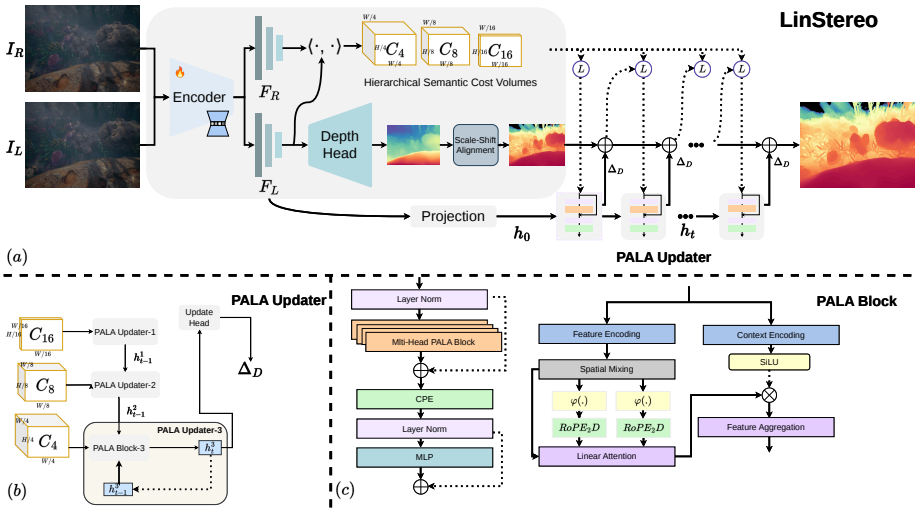


Fig. 1: Detailed Architecture of LinStereo: Our model replaces the ConvGRU update operator with the Position-Aware Linear Attention (PALA) updater, which enables global spatial reasoning over cost volume features at linear complexity by attention. The attention path applies kernel-activated queries and keys with 2D rotary positional encoding, while a parallel context modulation branch provides adaptive gating for selective feature aggregation.

3.2 Feature Extraction

Before describing the main architectural components, we introduce the shared feature extraction backbone. Given a rectified stereo pair $(\mathbf{I}_L, \mathbf{I}_R) \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the image height and width, we employ Depth Anything V3 (DA3) [25] as a shared backbone for both stereo feature extraction and monocular depth estimation. As the latest Depth Anything release, DA3 offers more detailed and geometrically consistent monocular depth than earlier versions, motivating its use here. It consists of a DINOv2 ViT-B/14 encoder [30] and a Dual-DPT head [32], which first reassembles multi-scale intermediate ViT representations through shared modules and then routes them to separate fusion branches for depth and ray prediction. We leverage the intermediate outputs of this architecture in two complementary ways from a single forward pass. For stereo correlation, we extract multi-scale feature representations after the shared reassembly layers but before fusion branches. This yields multi-scale features $\{\mathbf{F}_{L,s}, \mathbf{F}_{R,s}\}_{s\in\{4,8,16\}}$ with $\mathbf{F}_{L/R,s} \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times D_s}$, where s denotes the spatial downsampling factor. These features retain both fine geometric detail and high-level semantic structure from the VFM, and are subsequently projected to a uniform channel dimension c_f=128 for HSCV (Sec. 3.3). Meanwhile, the depth fusion branch produces a dense monocular depth map that provides the initialization signal for iterative refinement (Sec. 3.4), adding only the cost of the lightweight fusion layers.

3.3 Hierarchical Semantic Cost Volumes(HSCV)

Existing iterative methods [26, 45] build a single correlation volume at a fixed resolution (e.g., 1/4), restricting matching to a narrow semantic bandwidth and ignoring the hierarchy of modern VFM backbones. HSCV instead makes two contributions. First, *utilization*: we build a per-scale correlation volume from each VFM scale and keep the full set simultaneously accessible, so PALA can re-query any scale at every iteration, unlike one-shot multi-scale fusion [50] or coarse-to-fine cascades [14] that consume coarser volumes once and discard them. Second, *representation*: each per-scale volume carries its own internal disparity-axis pyramid for coarse-to-fine matching while preserving spatial resolution.

Given the stereo feature maps $\{\mathbf{F}_{L,s}, \mathbf{F}_{R,s}\}$ for each downsampling factor $s \in \{4, 8, 16\}$ (as defined in Sec. 3.2), we first project each through dedicated modules consisting of instance-normalized residual blocks and 1×1 convolutions to obtain $\hat{\mathbf{F}}_{L/R,s} = \psi_s(\mathbf{F}_{L/R,s})$. We then compute the 1D correlation volume along the epipolar line at each scale:

$$\mathbf{C}_s(i,j,d) = \frac{1}{\sqrt{C_f}} \left\langle \hat{\mathbf{F}}_{L,s}(i,j),\ \hat{\mathbf{F}}_{R,s}(i,j-d) \right\rangle, \tag{1}$$

where i, j index spatial positions and d indexes disparity displacement. To capture correspondences at both fine-grained and large displacements, each correlation volume $\mathbf{C}_s$ is organized into a 4-level pyramid by repeatedly downsampling along the disparity dimension, providing progressively larger search ranges while preserving spatial resolution.

This design introduces a two-level matching hierarchy: across semantic scales from the VFM backbone and within each scale's disparity pyramid. Each level of the iterative refinement thus receives semantically appropriate correlation signals at its own operating resolution, rather than relying solely on inter-scale hidden state propagation.

3.4 Depth Prior Initialization (DPI)

VFM-based monocular depth estimators capture rich relative structure with strong cross-domain generalization, yet their predictions are affine-invariant and lack the metric scale of a specific stereo rig. While metric monocular models [5, 19] exist, their absolute scale is a statistical prior that drifts substantially in out-of-distribution scenes (*e.g.*, underwater). Affine-invariant models avoid this commitment and preserve more robust relative structure across domains; since a calibrated stereo pair already provides geometric correspondences, the metric scale can be recovered directly from the data. The DA3 depth fusion branch produces an affine-invariant depth map $\hat{\mathbf{D}}_{\text{mono}}$ from the shared forward pass. To convert it into a metrically meaningful disparity initialization, we perform a lightweight scale-shift alignment, replacing the zero-disparity start used in [26] and its descendants that wastes early iterations on gross layout recovery.

Scale-Shift Alignment. The disparity is inversely proportional to depth ($\mathbf{d} \propto 1/\mathbf{D}$), therefore we convert the monocular prediction via $\mathbf{d} = \alpha/\hat{\mathbf{D}}_{\text{mono}} + \beta$, where α absorbs the unknown metric scale and β accounts for any additive offset. We estimate α and β from sparse correspondences: SIFT keypoints are extracted from the rectified pair and matched along epipolar lines, yielding references $\{(p_k, d_k^{\text{sp}})\}_{k=1}^{K}$. Compared to learning-based matchers [3, 43], SIFT is sufficient here since only sparse correspondences are needed. The parameters are obtained by least-squares:

$$(\alpha^*, \beta^*) = \arg\min_{\alpha,\beta} \sum_{k=1}^{K} \left\| d_k^{\text{sp}} - \left(\frac{\alpha}{\hat{\mathbf{D}}_{\text{mono}}(p_k)} + \beta \right) \right\|^2, \tag{2}$$

giving the initial disparity map:

$$\mathbf{d}^{(0)} = \frac{\alpha^*}{\hat{\mathbf{D}}_{\text{mono}}} + \beta^*. \tag{3}$$

When inlier matches fall below $K_{\min}$=20, we fall back to zero-disparity initialization (Fig. 1-(a)). We note that $\mathbf{d}^{(0)}$ serves as a coarse warm start; VFM depths do not strictly follow a global affine model, and handcrafted matches introduce localization noise. The iterative refinement with PALA (Sec. 3.5) is therefore essential to recover fine-grained disparity.

3.5 Position-Aware Linear Attention (PALA) Update

Existing recurrent update operators reason over a local spatial neighborhood per iteration, yet effective stereo refinement should propagate context across the entire spatial extent of the feature map, for instance, transferring reliable depth estimates from textured regions to nearby textureless areas. PALA achieves this through linear attention, which aggregates global spatial context with $\mathcal{O}(N)$ complexity per iteration, enabling each update step to attend to the full feature map without the quadratic cost of softmax attention. LinStereo employs three PALA updaters, one per scale, each containing a single PALA block that processes features at three resolution scales.

Multi-Scale PALA Updater As shown in Fig. 1-(b), at each refinement iteration t, given the current disparity estimate $\mathbf{d}_s^{(t)}$ for downsampling factor s, we perform a lookup into $\mathbf{C}_s$ around the predicted correspondence across all pyramid levels, and concatenate the retrieved values to form the per-scale correlation feature $\mathbf{c}_s^{(t)}$. A dedicated motion encoder then encodes the matching signal as $\mathbf{m}_s^{(t)} = \text{MotionEnc}_s(\mathbf{d}_s^{(t)}, \mathbf{c}_s^{(t)})$. Unlike prior work [26] where only the finest scale receives such inputs, our HSCV design supplies each scale with its own matching

signal, and hidden states $\{\mathbf{h}_s\}_{s\in\{4,8,16\}}$ are updated in coarse-to-fine order:

$$\mathbf{h}_{16}^{(t+1)} = \mathrm{PALA}_{16}\Big(\mathbf{h}_{16}^{(t)}, \mathbf{m}_{16}^{(t)}, \mathbf{h}_{8}^{(t)}\Big), \tag{4}$$

$$\mathbf{h}_{8}^{(t+1)} = \mathrm{PALA}_{8}\Big(\mathbf{h}_{8}^{(t)}, \mathbf{m}_{8}^{(t)}, \mathbf{h}_{4}^{(t)}, \mathbf{h}_{16}^{(t+1)}\Big), \tag{5}$$

$$\mathbf{h}_{4}^{(t+1)} = \mathrm{PALA}_{4}\Big(\mathbf{h}_{4}^{(t)}, \mathbf{m}_{4}^{(t)}, \mathbf{h}_{8}^{(t+1)}\Big), \tag{6}$$

where $\mathbf{h}_4^{(t+1)}$ produces the disparity update for final prediction.

PALA Block Architecture Standard linear attention achieves $\mathcal{O}(N)$ complexity by exploiting the associativity of matrix multiplication: rather than an $N\times N$ attention matrix, one precomputes the key-value product $\mathbf{K}^\top\mathbf{V} \in \mathbb{R}^{C_h\times C_h}$ and shares it across all query positions. However, this very associativity collapses all pairwise spatial relationships into a single global summary, making the resulting attention position-agnostic. For stereo depth refinement this is especially problematic: disparity updates are inherently spatially structured. For example, nearby pixels generally share similar depth, and discontinuities align with object boundaries at specific locations. A position-blind mechanism cannot distinguish these cases.

In the proposed PALA block (Fig. 1-*c*), we restore spatial awareness without sacrificing linear complexity by incorporating a relative position encoding through Rotary Position Embeddings (RoPE) [38]. Given kernel-mapped queries and keys $\hat{\mathbf{Q}} = \phi(\mathbf{W}_Q\mathbf{x})$, $\hat{\mathbf{K}} = \phi(\mathbf{W}_K\mathbf{x})$ with the non-negative kernel $\phi(\mathbf{z})$, we apply 2D RoPE extended to both height and width axes to obtain position-aware variants $\tilde{\mathbf{Q}} = \mathrm{RoPE}(\hat{\mathbf{Q}})$ and $\tilde{\mathbf{K}} = \mathrm{RoPE}(\hat{\mathbf{K}})$. The linear attention output is then:

$$\mathbf{O}_{\mathrm{attn}} = \tilde{\mathbf{Q}}\Big(\tilde{\mathbf{K}}^\top\mathbf{V}\Big)\cdot\left(\hat{\mathbf{Q}}\,\overline{\hat{\mathbf{K}}}^{\top} + \epsilon\right)^{-1}, \tag{7}$$

where $\overline{\hat{\mathbf{K}}}=\frac{1}{N}\sum_j \hat{\mathbf{K}}_j$. Crucially, RoPE is appliedasymmetrically: only the numerator uses position-augmented $\tilde{\mathbf{Q}}, \tilde{\mathbf{K}}$ to encode relative spatial relationships, while the denominator retains the original $\hat{\mathbf{Q}}, \hat{\mathbf{K}}$ for stable normalization. Applying RoPE uniformly would make the normalizing factor position-dependent, distorting attention magnitudes and destabilizing training. This preserves $\mathcal{O}(N\cdot C_h^2)$ complexity, while restoring the spatial structure that vanilla linear attention discards. Beyond the relative encoding within the attention mechanism, stereo disparity fields are locally smooth: neighboring pixels on the same surface share similar depth. We therefore complement the attention with local spatial encoding on the value branch to reinforce this neighborhood coherence, and absolute position encoding flanking the block to anchor features to their spatial coordinates, which helps the network distinguish systematic depth variation across different image regions.

Since PALA operates within an iterative refinement loop, we employ a gated update to control information flow across iterations:

$$\mathbf{z}_g = \sigma(\mathrm{Conv}_{3\times3}([\mathbf{h}\|\mathbf{O}])), \quad \mathbf{h}_{\mathrm{new}} = (1-\mathbf{z}_g)\odot\mathbf{h} + \mathbf{z}_g\odot\tanh(\mathbf{O}). \tag{8}$$

Here, $\|$ denotes channel-wise concatenation, $\sigma(\cdot)$ the sigmoid function, and $\odot$ element-wise multiplication. In stereo matching, correlation evidence is inherently non-uniform: textured regions yield strong, unambiguous matches, while occluded or textureless areas produce noisy signals. The gate $\mathbf{z}_g$ enables each location to aggressively incorporate new evidence where matching is confident and to preserve the accumulated estimate where it is not, allowing reliable disparity to gradually propagate into uncertain regions across iterations.

3.6 Loss and Training Strategy

The finest-scale hidden state $\mathbf{h}_4^{(t+1)}$ is decoded into a disparity residual $\Delta\mathbf{d}^{(t)}$ through stacked depth-wise convolutional blocks. The training objective supervises each iterative refinement step with exponentially increasing weights:

$$\mathcal{L} = \sum_{t=1}^{T} \gamma^{T-t} \Big(\Big\| \mathbf{d}^{(t)} - \mathbf{d}_{\text{gt}} \Big\|_1 + \lambda_s \mathcal{L}_{\text{smooth}}(\mathbf{d}^{(t)}) + \lambda_g \mathcal{L}_{\text{grad}}(\mathbf{d}^{(t)}) \Big), \tag{9}$$

where γ assigns progressively higher weight to later iterations. $\mathcal{L}_{\text{smooth}}$ is an edge-aware smoothness term that penalizes disparity gradients in regions with low image intensity variation, encouraging piece-wise smooth predictions while preserving discontinuities at object boundaries. $\mathcal{L}_{\text{grad}}$ [33] computes ℓ_1 differences between predicted and ground-truth disparity gradients at multiple spatial scales, guiding the network to recover accurate depth edges. Both auxiliary terms carry small weights (λ_s, λ_g). Allh hyperparameters are provided in Sec. 4.2.

4 Experiments

4.1 Datasets and Metrics

Training. LinStereo is trained exclusively on **SceneFlow** [28], a large-scale synthetic dataset comprising approximately 35,000 stereo pairs with dense disparity ground truth. No real-world, underwater, or domain-specific data is used at any stage of training. *Standard Benchmarks.* Cross-domain generalization is evaluated on five standard stereo benchmarks: **KITTI 2015** [29] and **KITTI 2012** [13], which provide real-world driving scenes; **Middlebury (H)** [34], an indoor benchmark at half-resolution; **ETH3D** [35], a high-precision LiDAR-aligned dataset; and **Booster (Q)** [31], evaluated at quarter resolution and notable for containing transparent and specular surfaces. Following standard protocol, end-point error (EPE) and percentage of pixels with disparity error $> k$ pixels (bad-k) are reported. *Underwater Benchmarks.* To assess generalization under photometric degradation, two underwater benchmarks are adopted. **TartanAir-UW** consists of 13,583 stereo pairs from 22 underwater sequences in TartanAir [40], representing simulated subsea scenes with realistic light scattering. **SQUID** [4] provides 57 real-world underwater stereo pairs across 4 different scenes, covering a range of water visibility conditions. Both benchmarks operate at moderate-to-large depth scales. Since these benchmarks provide depth

rather than disparity ground truth, depth-domain metrics are adopted following prior work [3]: AbsRel, SqRel, RMSE, LogRMSE, and accuracy thresholds $\delta < 1.25^k$ ($k = 1, 2, 3$). Additionally, we contribute a synthetic underwater stereo corpus (**SeaStereo-Dataset**, ∼40K stereo pairs) with dense disparity ground truth under realistic subsea optical conditions; dataset details and evaluation are provided in the supplementary material. SeaStereo is built by rendering ShapeNetCore [6] foreground objects (arbitrary `.obj` meshes) over 3D marine backgrounds (coral, fish, and shipwrecks) under varying Jerlov water types.
Real-World Evaluation. LinStereo is further validated on a controlled **laboratory tank** dataset collected in-house at close range ($< 2\,\mathrm{m}$). In contrast to TartanAir-UW and SQUID, which evaluate at moderate-to-large depth scales, this benchmark tests near-range precision under genuine underwater conditions, where small absolute errors translate to large relative depth deviations. Ground-truth depth is obtained via AprilTag (16h5) markers with PnP-based pose recovery; detailed experimental setup is described in the supplementary material.

4.2 Implementation Details

The DA3 backbone is kept **fully frozen**; only the decoder is trained. For DPI, SIFT matching uses a minimum inlier threshold $K_{\min} = 20$, below which we fall back to zero-disparity initialization; this occurs for 0% of TartanAir-UW and 3.7% of SQUID frames and costs only $+0.08\,\mathrm{px}$ EPE. We train with AdamW for 200K steps on SceneFlow (batch size 8, 320×640 crops, one NVIDIA H100), with a one-cycle schedule peaking at 2×10^{-4}. The loss (Eq. (9)) spans $T = 8$ refinement iterations ($\gamma = 0.9$) with smoothness ($\lambda_s = 0.001$) and gradient ($\lambda_g = 0.01$) terms. For real-time use, the budget can drop to $T = 2$ (supplementary).

4.3 Main Results

Standard-Benchmark Generalization. Tab. 1 reports *zero-shot* cross-domain results. Each method uses its official released weights (trained on SceneFlow unless marked ✓) and is evaluated on the training splits with public ground truth (Middlebury at half resolution (H)). FoundationStereo [43] (ViT-L, with substantial extra training data) is the strongest baseline; its larger backbone and extra data account for its lead on these standard benchmarks. We deliberately keep LinStereo a ViT-B[1], SceneFlow-only model to isolate the decoder's contribution. Even in this smaller-backbone, same-data setting, it remains highly competitive: it matches or beats the ViT-L BridgeDepth [15] on 13/14 metrics (trailing only on ETH3D bad-1) and outperforms MonSter [9] on 9/14. The gains thus stem from the proposed architectural components rather than backbone capacity or extra data. The most pronounced improvement appears on occluded regions of Middlebury (H), where EPE is 37% lower than the previous best [20]. This validates PALA's global context propagation: reliable disparity estimates from well-matched visible areas are propagated to occluded pixels

[1] ViT-B keeps training feasible on modest hardware; our design is backbone-agnostic.

Table 1: Cross-domain generalization on standard stereo benchmarks. Booster (Q): quarter-resolution evaluation. Extra Data (✓): method uses additional training data beyond SceneFlow [28]. **Bold**/underline: best/second-best among methods without extra data.

Method	Extra Data	KITTI 2015		KITTI 2012		Middlebury (H)						ETH3D		Booster (Q)	
						All		NonOcc		Occ					
		EPE	bad3	EPE	bad3	EPE	bad2	EPE	bad2	EPE	bad2	EPE	bad1	EPE	bad2
PSMNet [7]	—	4.05	28.51	3.77	27.34	11.48	36.06	10.56	32.11	17.69	62.58	2.15	15.13	14.17	49.79
RAFT-Stereo [26]	—	1.13	5.69	0.90	4.35	1.92	12.6	1.09	8.65	3.31	26.39	0.36	3.3	4.18	17.64
UniMatch [49]	—	1.63	11.01	1.71	12.04	4.21	34.72	3.96	31.52	5.70	56.51	0.69	18.61	8.57	47.30
MoCha-Stereo [8]	—	1.29	5.97	1.02	4.83	2.66	10.18	2.49	7.96	3.84	24.16	0.28	3.47	3.88	16.82
NMRF [16]	—	1.17	5.31	0.92	4.63	2.91	13.36	2.73	10.90	4.22	29.27	0.31	3.8	5.05	26.22
MGStereo [53]	—	1.13	5.62	0.87	4.16	1.15	8.39	0.85	5.67	2.89	26.50	0.25	1.88	2.26	11.02
Stereo Anywhere(ViT-L) [3]	—	1.07	4.93	0.83	3.90	0.94	6.96	**0.79**	**4.75**	2.67	20.34	0.24	2.66	2.21	**9.91**
DEFOM-Stereo (ViT-L) [20]	—	1.06	4.99	0.84	3.76	0.84	**5.91**	0.90	5.76	2.11	19.34	0.35	2.35	3.52	13.20
MonSter (ViT-L) [9]	—	**0.89**	**3.58**	**0.75**	3.58	1.98	10.41	2.03	8.79	2.62	21.56	**0.23**	1.41	3.08	17.45
BridgeDepth (ViT-L) [15]	—	1.13	4.90	0.86	4.39	2.21	10.27	2.25	8.66	2.68	20.42	0.24	**1.33**	4.71	18.79
Ours(ViT-B)	—	1.01	4.54	0.76	**3.49**	**0.83**	6.01	0.89	5.67	**1.33**	**8.69**	0.24	2.41	**2.14**	9.93
IGEV-Stereo [45]	✓	1.21	6.03	1.03	5.13	2.63	11.93	2.27	9.49	5.02	26.04	0.33	4.0	4.26	17.58
Selective-RAFT [41]	✓	1.27	6.68	1.08	5.19	2.34	12.04	2.05	9.45	4.17	27.4	0.34	4.36	4.14	19.52
Selective-IGEV [41]	✓	1.25	6.06	1.08	5.64	2.59	11.79	2.31	9.22	4.35	28.10	0.33	4.05	4.62	19.28
IGEV++ [46]	✓	1.27	6.22	1.20	6.81	3.21	10.85	2.41	7.86	6.83	28.68	0.35	4.60	5.00	18.62
FoundationStereo (ViT-L) [43]	✓	0.95	3.65	0.71	3.18	0.71	3.47	0.42	1.46	2.67	16.44	0.21	1.14	1.77	6.89

that lack direct photometric evidence, which local-window updaters struggle to achieve. Booster (Q) is the most challenging benchmark, with transparent and specular surfaces that violate photometric consistency; even here, LinStereo remains stable, trailing the best by only a marginal Bad-2 gap (9.93 vs. 9.91 [3]) that lies within statistical variation.

Underwater Generalization. Underwater scenes are a stringent cross-domain test, with photometric degradation from scattering and attenuation. On TartanAir-UW and SQUID (Tab. 2, $T=8$), LinStereo, trained only on SceneFlow, achieves the best results on both, surpassing even methods trained on real-world or domain-specific data [41, 43, 45, 46]. The largest gains track each benchmark's dominant degradation: 31% lower RMSE than FoundationStereo [43] on TartanAir-UW (long-range backscatter) and 26% lower AbsRel than IGEV++ [46] on SQUID (color attenuation), indicating PALA propagates reliable estimates to the most degraded regions.

Qualitative Evaluation. Fig. 2a compares predictions on Booster (Q) and ETH3D. Recent SOTA methods produce visually near-identical depth maps on these well-conditioned scenes; LinStereo shows sharper thin structures and handles non-Lambertian surfaces (*e.g.*, transparent bottles and car windows), consistent with its quantitative lead. This visual near-parity motivates our underwater evaluation, where photometric degradation amplifies inter-method differences. As shown in Fig. 2b, competing methods tend to overestimate depth in distant regions and produce inconsistent near-field estimates, whereas LinStereo preserves accurate depth scale at both ranges.

Table 2: Quantitative comparison on underwater depth estimation. TartanAir-UW: underwater images from TartanAir [40]. SQUID: real-world underwater dataset [4]. ✓: uses extra training data beyond SceneFlow [28].

Method	Extra Data	TartanAir UW							SQUID						
		AbsRel↓	SqRel↓	RMSE↓	LogRMSE↓	A1↑	A2↑	A3↑	AbsRel↓	SqRel↓	RMSE↓	LogRMSE↓	A1↑	A2↑	A3↑
UniMatch [49]	—	0.17	2.59	6.35	0.26	0.759	0.830	0.837	1.38	20.91	6.42	0.66	0.561	0.660	0.734
MoCha-Stereo [8]	—	0.15	2.30	6.02	0.25	0.769	0.835	0.843	0.15	0.61	1.55	0.20	0.875	0.931	0.958
NMRF [16]	—	0.11	1.08	4.63	0.21	0.811	0.856	0.870	0.43	5.27	2.57	0.33	0.844	0.904	0.935
PSMNet [7]	—	0.10	1.02	4.56	0.21	0.813	0.857	0.871	0.46	4.31	3.43	0.51	0.736	0.813	0.853
RAFT-Stereo [26]	—	0.08	0.65	4.36	0.20	0.902	0.963	0.984	0.07	0.27	1.25	0.12	0.937	0.971	0.987
MGStereo [53]	—	0.08	0.55	3.69	0.19	0.911	0.968	0.987	0.09	0.71	1.99	0.16	0.925	0.958	0.976
Stereo Anywhere [3]	—	0.06	0.41	3.24	0.18	0.946	0.980	0.989	0.07	0.40	1.46	0.13	0.937	0.976	0.985
DEFOM-Stereo [20]	—	0.05	0.41	3.13	0.17	0.953	0.982	0.990	0.09	0.63	2.00	0.16	0.915	0.954	0.977
Ours	—	**0.04**	**0.19**	**2.08**	**0.09**	**0.959**	**0.992**	**0.998**	**0.04**	**0.12**	**0.90**	**0.08**	**0.970**	**0.990**	**0.996**
Selective-Stereo (RAFT) [41]	✓	0.09	0.67	4.31	0.21	0.902	0.962	0.981	0.11	0.22	1.16	0.16	0.877	0.940	0.968
Selective-Stereo (IGEV) [41]	✓	0.11	1.02	5.01	0.23	0.856	0.942	0.976	0.08	0.21	1.05	0.14	0.932	0.966	0.980
IGEV-Stereo [45]	✓	0.10	0.90	4.68	0.21	0.891	0.955	0.979	0.20	1.35	2.68	0.46	0.760	0.821	0.863
IGEV++ [46]	✓	0.09	0.81	4.37	0.20	0.905	0.962	0.983	0.06	0.22	1.11	0.12	0.950	0.981	0.990
FoundationStereo [43]	✓	0.05	0.40	3.01	0.16	**0.959**	0.984	0.991	0.07	0.30	1.36	0.13	0.940	0.976	0.987

Table 3: Quantitative comparison on close-range laboratory tank images ($< 2\,\mathrm{m}$ depth). Extra Data (✓) indicates training data beyond SceneFlow [28]. Accuracy thresholds rounded to two decimal places; full-precision results in supplimentary materials. Best results in **bold**.

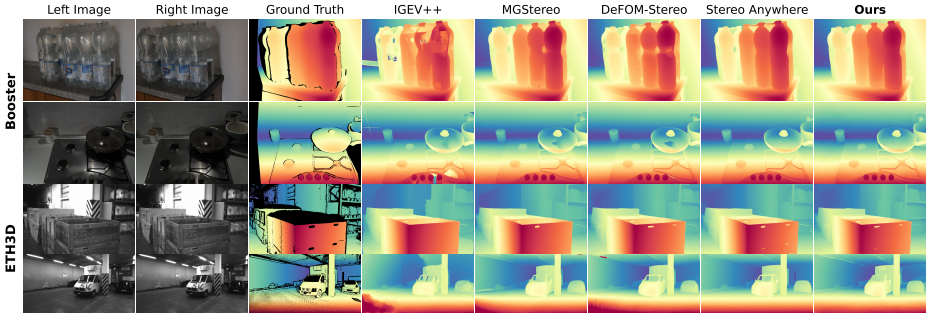
Method	Extra Data	Laboratory Tank						
		AbsRel↓	SqRel↓	RMSE↓	LogRMSE↓	A1↑	A2↑	A3↑
PSMNet [7]	—	0.18	0.16	0.34	0.28	0.87	0.90	0.93
UniMatch [49]	—	0.08	0.04	0.17	0.16	0.93	0.96	0.98
Stereo Anywhere [3]	—	0.09	0.06	0.20	0.19	0.92	0.93	0.97
RAFT-Stereo [26]	—	0.06	0.03	0.15	0.14	0.94	0.96	0.99
MGStereo [53]	—	0.08	0.05	0.18	0.17	0.93	0.94	0.98
Ours (T=8)	—	**0.04**	**0.01**	**0.07**	**0.07**	**0.98**	**0.99**	**1.00**
Selective-Stereo (RAFT) [41]	✓	0.07	0.03	0.14	0.14	0.94	0.97	0.99
Selective-Stereo (IGEV) [41]	✓	0.08	0.04	0.16	0.16	0.91	0.95	0.99
IGEV-Stereo [45]	✓	0.06	0.02	0.13	0.12	0.94	0.97	0.99
IGEV++ [46]	✓	0.05	0.02	0.12	0.12	0.96	0.97	0.99
FoundationStereo [43]	✓	0.07	0.04	0.18	0.17	0.93	0.94	0.98

4.4 Real-World Evaluation

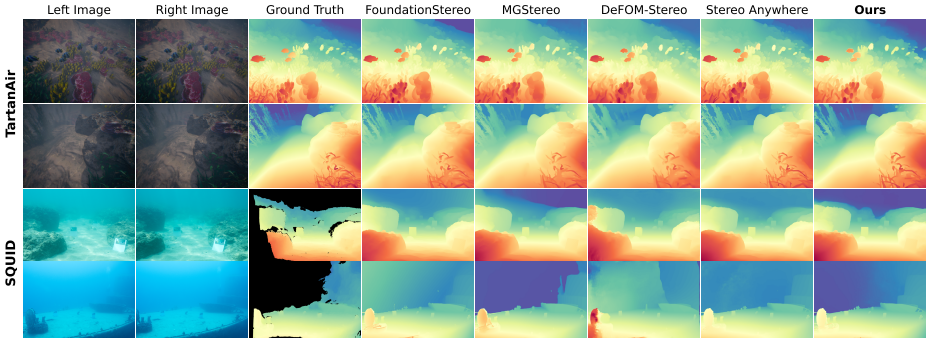
LinStereo is additionally evaluated on a controlled laboratory water tank at close range ($< 2\,\mathrm{m}$) with AprilTag-based ground truth (Sec. 4.1). As shown in Tab. 3, LinStereo outperforms all compared methods by a substantial margin, including those leveraging multiple additional training datasets. We note that the reference depth comes from a CAD model of the rig (frame and ropes), localized in each view via AprilTags; only these valid-GT pixels are scored, so the larger background errors produced by some baselines fall outside this quantitative evaluation. The result confirms that the proposed architecture generalizes robustly even under extreme near-range, real-world underwater conditions. Full-precision table and qualitative results are provided in supplementary materials.

4.5 Ablation Study

All ablations are evaluated on KITTI 2015 and TartanAir-UW to capture the model's behavior on different domains. *PALA Design Ablation.* Tab. 4 traces a



(a) Booster(Q) and ETH3D benchmarks. Black regions in ground truth denote unavailable disparity.



(b) Underwater benchmarks. Black regions in SQUID ground truth denote unavailable depth.

Fig. 2: Qualitative comparison on standard and underwater stereo benchmarks.

progressive build-up of the PALA block from a position-agnostic linear attention baseline (A). Each addition yields consistent gains on both benchmarks, with global spatial encoding (A→B) providing the largest single improvement, confirming that restoring positional structure is the most critical enhancement to vanilla linear attention. Within this positional encoding, an asymmetric RoPE further outperforms its symmetric counterpart (Tab. 4). Local spatial encoding and adaptive gating contribute complementary refinements, and the full PALA configuration (D) achieves the best overall performance.

Component Ablation. Tab. 5 reports the results for all combinations of the three contributions. Each component individually improves over the baseline, with PALA contributing the largest single gain, consistent with its role as the primary architectural change. Notably, the full combination yields super-additive improvements. This is because PALA's global reasoning benefits from the richer per-scale correlations that HSCV provides, while DPI's warm start allows convergence within fewer iterations. Among pairwise combinations, HSCV + PALA achieves the second best performance (EPE 1.10), suggesting that global attention benefits most from multi-scale matching signals. Replacing the hierarchical design with a single pooled multi-scale feature (HSCV pooled) degrades accuracy, confirming that scale-aligned correlation better exploits VFM representations.

Table 4: PALA design ablation: progressive addition of components to global linear attention (A). (B) adds global spatial encoding; (C) further adds local spatial encoding; (D) adds adaptive gating to obtain Full PALA. Evaluated on KITTI 2015 and TartanAir-UW ($T=8$). Best in **bold**.

Experiment	KITTI 2015		TartanAir-UW	
	EPE↓	bad3↓	AbsRel↓	RMSE↓
(A) Global linear attention	1.18	5.32	0.052	2.55
(B) (A) + Global spatial encoding	1.12	5.05	0.047	2.38
(C) (B) + Local spatial encoding	1.06	4.76	0.043	2.22
(D) (C) + Adaptive gating **[Full PALA]**	**1.01**	**4.54**	**0.04**	**2.08**
with symmetric RoPE	1.05	4.76	0.042	2.18

Table 5: Ablation studies. All models are trained on SceneFlow only and evaluated on KITTI 2015 and TartanAir-UW. Best results in **bold**. In (a), ✓ = enabled; "HSCV (pooled)" uses a single pooled multi-scale feature.

(a) Component ablation

HSCV	DPI	PALA	KITTI 2015		TartanAir-UW	
			EPE↓	bad3↓	AbsRel↓	RMSE↓
			1.42	6.58	0.068	3.21
✓			1.33	6.14	0.061	2.91
	✓		1.32	6.02	0.059	2.85
		✓	1.26	5.78	0.055	2.64
✓	✓		1.21	5.51	0.052	2.52
✓		✓	1.10	4.95	0.045	2.28
	✓	✓	1.13	5.12	0.047	2.35
✓	✓	✓	**1.01**	**4.54**	**0.04**	**2.08**
pooled	✓	✓	1.07	4.82	0.043	2.21

(b) Training hyperparameters

LR	BS	T	KITTI 2015		TartanAir-UW	
			EPE↓	bad3↓	AbsRel↓	RMSE↓
1×10^{-4}	8	8	1.09	4.88	0.045	2.27
$\mathbf{2\times10^{-4}}$	**8**	**8**	**1.01**	**4.54**	**0.04**	**2.08**
4×10^{-4}	8	8	1.16	5.22	0.050	2.45
2×10^{-4}	4	8	1.07	4.79	0.044	2.22
2×10^{-4}	16	8	1.05	4.68	0.042	2.16
2×10^{-4}	8	2	1.24	5.65	0.055	2.62
2×10^{-4}	8	4	1.10	4.93	0.046	2.30
2×10^{-4}	8	12	1.03	4.61	0.041	2.12
2×10^{-4}	8	16	1.02	4.58	0.041	2.10

Training Hyperparameters. Tab. 5(b) sweeps peak learning rate, batch size, and refinement iterations. Performance is relatively robust across the tested ranges, with the optimal setting (LR $= 2\times10^{-4}$, BS $= 8$, $T = 8$) achieving the best accuracy. The most notable finding is that $T=12$ yields no improvement over $T=8$, indicating diminishing returns beyond 8 iterations.

Inference Efficiency. LinStereo's iterative design lets the refinement budget shrink at test time: at $T=2$ it runs at 80 ms per frame (12.5 FPS at 480×640 on a single RTX 4500) while still surpassing every efficient stereo baseline on both underwater benchmarks, most strikingly reaching AbsRel 0.05 on SQUID where the strongest competitor stays at ≥ 0.71 (Tab. 6; the full efficient-method comparison is given in the supplementary material). Of the 127 M parameters, ∼120 M (94%) lie in the frozen DA3 backbone, while the proposed PALA, HSCV, and DPI together add only ∼7 M; per refinement step PALA's global update is no costlier than a local ConvGRU operator (Tab. 7), so linear attention supplies global reasoning without the quadratic cost of softmax attention.

Table 6: Computational efficiency comparison. Measured at 480×640 resolution on a single NVIDIA RTX 4500 GPU (24 GB).

Method	Params (M)↓	GFLOPs↓	Time (ms)↓	FPS↑
LightStereo-S [17]	3.44	**45.4**	**9.9**	**101.0**
CoEx [2]	2.73	70.6	11.0	91.0
CGI-Stereo [48]	3.50	81.6	12.8	78.1
Fast-ACVNet [47]	3.08	104.6	13.2	76.0
ADStereo [42]	7.10	201.6	15.4	64.7
RAFT-Stereo† [26]	9.87	296.4	18.0	55.7
RT-IGEV [46]	4.17	354.2	24.7	40.5
MobileStereoNet [36]	**2.35**	175.6	37.7	26.5
AANet [50]	3.93	–	65.8	15.2
Ours	127.0	770.4	80	12.5

Table 7: Per-iteration update-operator latency. Measured at 480×640 resolution on a single NVIDIA RTX 4500 GPU (24 GB); $\pm$ denotes the 95% confidence interval.

Update operator	Latency (ms)↓
PALA (Ours)	3.50 ± 0.05
RAFT-Stereo ConvGRU [26]	3.63 ± 0.06
IGEV ConvGRU [45]	3.43 ± 0.03

5 Conclusions, Limitations and Future Work

We proposed LinStereo, an iterative stereo depth estimation framework that addresses the information flow between VFM backbones and the iterative update through three complementary mechanisms: PALA for global context aggregation at linear complexity, HSCV for scale-aligned semantic matching, and DPI for geometrically informed warm start. Trained exclusively on SceneFlow, LinStereo achieves state-of-the-art-level accuracy on standard stereo benchmarks and demonstrates strong cross-domain generalization to underwater environments, achieving the best overall accuracy on both simulated and real-world underwater benchmarks.

Several directions remain for future work. LinStereo builds upon a frozen DA3 backbone, which accounts for most of the inference cost and can limit deployment on resource-constrained platforms at the full iteration budget. Distilling the backbone into a lightweight architecture could reduce latency while preserving cross-domain feature quality. In addition, the DPI module currently relies on handcrafted SIFT features for scale–shift alignment. Replacing this component with a lightweight learned sparse matcher may provide more robust initialization and improved adaptability across domains in the feature.

Acknowledgements

We thank Jay Zhang, Dr. Gideon Billings for collecting the underwater dataset. This research was funded by the Australian Government through the Australian Research Council (ARC) Research Hub for Intelligent Robotic Systems for Real-Time Asset Management (IH210100030).

References

1. Akkaynak, D., Treibitz, T.: A revised underwater image formation model. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6723–6732. IEEE (2018). https://doi.org/10.1109/CVPR.2018.00703
2. Bangunharcana, A., Cho, J.W., Lee, S., Kweon, I.S., Kim, K.S., Kim, S.: Correlate-and-excite: Real-time stereo matching via guided cost volume excitation. In: IEEE/RJS International Conference on Intelligent RObots and Systems. pp. 3542–3548. IEEE, IEEE (2021). https://doi.org/10.1109/IROS51168.2021.9635909
3. Bartolomei, L., Tosi, F., Poggi, M., Mattoccia, S.: Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail. In: Computer Vision and Pattern Recognition. pp. 1013–1027 (2024). https://doi.org/10.1109/CVPR52734.2025.00103
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 (2018). https://doi.org/10.1109/tpami.2020.2977624
5. Bochkovskiy, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: arXiv.org (2024). https://doi.org/10.48550/arXiv.2410.02073
6. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q.X., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: ShapeNet: An Information-Rich 3D Model Repository. Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago (2015)
7. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5410–5418. IEEE (2018). https://doi.org/10.1109/CVPR.2018.00567
8. Chen, Z., Long, W., Yao, H., Zhang, Y., Wang, B., Qin, Y., Wu, J.: Mocha-stereo: Motif channel attention network for stereo matching. In: Computer Vision and Pattern Recognition. pp. 27768–27777. IEEE (June 2024). https://doi.org/10.1109/CVPR52733.2024.02623
9. 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: Computer Vision and Pattern Recognition. pp. 6273–6282. IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.00588
10. Du, S., Zou, Y., Li, J., Liu, M., Li, Y., Shang, C., Shen, Q.: Pansharpening for thin-cloud contaminated remote sensing images: a unified framework and benchmark dataset. In: AAAI Conference on Artificial Intelligence. vol. 40, pp. 3696–3704. Association for the Advancement of Artificial Intelligence (AAAI) (2026). https://doi.org/10.1609/aaai.v40i5.37369
11. Du, S., Zou, Y., Wang, Z., Li, X., Li, Y., Shang, C., Shen, Q.: Unsupervised hyperspectral image super-resolution via self-supervised modality decoupling. International Journal of Computer Vision **134**, 152 (2024). https://doi.org/10.1007/s11263-026-02757-8
12. Duggal, S., Wang, S., Ma, W.C., Hu, R., Urtasun, R.: Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In: IEEE International Conference on Computer Vision. pp. 4383–4392. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00448
13. 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). https://doi.org/10.1109/CVPR.2012.6248074
14. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Computer Vision and Pattern Recognition. pp. 2495–2504 (2019). https://doi.org/10.1109/CVPR42600.2020.00257
15. Guan, T., Guo, J., Wang, C., Liu, Y.: Bridgedepth: Bridging monocular and stereo reasoning with latent alignment. In: IEEE International Conference on Computer Vision. pp. 27681–27691. IEEE (2025). https://doi.org/10.1109/ICCV51701.2025.02570
16. Guan, T., Wang, C., Liu, Y.H.: Neural markov random field for stereo matching. In: Computer Vision and Pattern Recognition. pp. 5459–5469. IEEE (June 2024). https://doi.org/10.1109/CVPR52733.2024.00522
17. Guo, X., Zhang, C., Nie, D., Zheng, W., Zhang, Y., Chen, L.: Lightstereo: Channel boost is all you need for efficient 2d cost aggregation. In: IEEE International Conference on Robotics and Automation (2024). https://doi.org/10.1109/ICRA55743.2025.11127711
18. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: Computer Vision and Pattern Recognition. pp. 3268–3277. IEEE (2019). https://doi.org/10.1109/CVPR.2019.00339
19. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10579–10596 (2024). https://doi.org/10.1109/TPAMI.2024.3444912
20. Jiang, H., Lou, Z., Ding, L., Xu, R., Tan, M., Jiang, W., Huang, R.: DEFOM-Stereo: Depth foundation model based stereo matching. In: Computer Vision and Pattern Recognition. pp. 21857–21867. IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.02036
21. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P.: End-to-end learning of geometry and context for deep stereo regression. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 66–75. IEEE (2017). https://doi.org/10.1109/iccv.2017.17
22. Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J., Izadi, S.: Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In: European Conference on Computer Vision. pp. 596–613. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01267-0_35
23. Li, J., Wang, P., Xiong, P., Cai, T., Yan, Z., Yang, L., Liu, J., Fan, H., Liu, S.: Practical stereo matching via cascaded recurrent network with adaptive correlation. In: Computer Vision and Pattern Recognition. pp. 16242–16251. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01578
24. Li, P., Zhang, Z., Holtz, D., Yu, H., Yang, Y., Lai, Y., Song, R., Geiger, A., Zell, A.: Spacedrive: Infusing spatial awareness into vlm-based autonomous driving. In: arXiv.org. pp. 40096–40107 (2025). https://doi.org/10.48550/arXiv.2512.10719
25. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv.org (2025). https://doi.org/10.48550/arXiv.2511.10647
26. Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: International Conference on 3D Vision. pp. 218–227. IEEE, IEEE (2021). https://doi.org/10.1109/3DV53792.2021.00032

27. Lv, Q., Dong, J., Li, Y., Chen, S., Yu, H., Zhang, S., Wang, W.: Uwstereo: A large synthetic dataset for underwater stereo matching. IEEE transactions on circuits and systems for video technology (Print) (2024). https://doi.org/10.1109/TCSVT.2025.3572044
28. 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: Computer Vision and Pattern Recognition. pp. 4040–4048 (2015). https://doi.org/10.1109/CVPR.2016.438
29. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Computer Vision and Pattern Recognition. pp. 3061–3070. IEEE (2015). https://doi.org/10.1109/CVPR.2015.7298925
30. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. (2023). https://doi.org/10.48550/arXiv.2304.07193
31. Ramirez, P.Z., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Stefano, L.D.: Open challenges in deep stereo: the booster dataset. In: Computer Vision and Pattern Recognition. pp. 21136–21146. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.02049
32. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: IEEE International Conference on Computer Vision. pp. 12159–12168. IEEE (October 2021). https://doi.org/10.1109/ICCV48922.2021.01196
33. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3), 1623–1637 (2019). https://doi.org/10.1109/TPAMI.2020.3019967
34. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nesic, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: German Conference on Pattern Recognition. pp. 31–42. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-11752-2_3
35. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Computer Vision and Pattern Recognition. pp. 2538–2547. IEEE (2017). https://doi.org/10.1109/CVPR.2017.272
36. Shamsafar, F., Woerz, S., Rahim, R., Zell, A.: Mobilestereonet: Towards lightweight deep networks for stereo matching. In: IEEE Workshop/Winter Conference on Applications of Computer Vision. pp. 2417–2426 (2021). https://doi.org/10.1109/WACV51458.2022.00075
37. Shen, Z., Dai, Y., Song, X., Rao, Z., Zhou, D., Zhang, L.: Pcw-net: Pyramid combination and warping cost volume for stereo matching. In: European Conference on Computer Vision. pp. 280–297. Springer (2020). https://doi.org/10.1007/978-3-031-19824-3_17
38. Su, J., Lu, Y., Pan, S., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2021). https://doi.org/10.1016/j.neucom.2023.127063
39. Tankovich, V., Häne, C., Fanello, S., Zhang, Y., Izadi, S., Bouaziz, S.: Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In: Computer Vision and Pattern Recognition. pp. 14362–14372 (2020). https://doi.org/10.1109/CVPR46437.2021.01413

40. 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: IEEE/RJS International Conference on Intelligent RObots and Systems. pp. 4909–4916. IEEE (2020). https://doi.org/10.1109/IROS45743.2020.9341801
41. Wang, X., Xu, G., Jia, H., Yang, X.: Selective-stereo: Adaptive frequency information selection for stereo matching. In: Computer Vision and Pattern Recognition. pp. 19701–19710. IEEE (June 2024). https://doi.org/10.1109/CVPR52733.2024.01863
42. Wang, Y., Li, K., Wang, L., Hu, J., Wu, D.O., Guo, Y.: Adstereo: Efficient stereo matching with adaptive downsampling and disparity alignment. IEEE Transactions on Image Processing **34**, 1204–1218 (2025). https://doi.org/10.1109/TIP.2025.3540282
43. Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: FoundationStereo: Zero-shot stereo matching. In: Computer Vision and Pattern Recognition. pp. 5249–5260. IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.00495
44. Xu, B., Xu, Y., Yang, X., Jia, W., Guo, Y.: Bilateral grid learning for stereo matching networks. In: Computer Vision and Pattern Recognition. pp. 12492–12501. IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.01231
45. Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. In: Computer Vision and Pattern Recognition. pp. 21919–21928. IEEE (June 2023). https://doi.org/10.1109/CVPR52729.2023.02099
46. 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 **47**, 7108–7122 (2024). https://doi.org/10.1109/TPAMI.2025.3569218
47. Xu, G., Wang, Y., Cheng, J., Tang, J., Yang, X.: Accurate and efficient stereo matching via attention concatenation volume. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(4), 2461–2474 (2022). https://doi.org/10.1109/TPAMI.2023.3335480
48. Xu, G., Zhou, H., Yang, X.: Cgi-stereo: Accurate and real-time stereo matching via context and geometry interaction. arXiv.org (2023). https://doi.org/10.48550/arXiv.2301.02789
49. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Yu, F., Tao, D., Geiger, A.: Unifying flow, stereo and depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 13941–13960 (2022). https://doi.org/10.1109/TPAMI.2023.3298645
50. Xu, H., Zhang, J.: Aanet: Adaptive aggregation network for efficient stereo matching. In: Computer Vision and Pattern Recognition. pp. 1956–1965. IEEE (2020). https://doi.org/10.1109/cvpr42600.2020.00203
51. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Computer Vision and Pattern Recognition. pp. 10371–10381. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.00987
52. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything V2. In: Neural Information Processing Systems. pp. 21875–21911. Neural Information Processing Systems Foundation, Inc. (NeurIPS) (2024). https://doi.org/10.48550/arXiv.2406.09414
53. Yao, C., Yu, L., Liu, Z., Zeng, J., Wu, Y., Jia, Y.: Diving into the fusion of monocular priors for generalized stereo matching. In: IEEE International Conference on Computer Vision. pp. 14887–14897. IEEE (2025). https://doi.org/10.1109/ICCV51701.2025.01381

54. Yu, H., Jin, Y., Canevaro, A., Schmidt, J., Jordan, J., Li, P., Kaufeld, M., Lindner, S., Betz, J., Stork, W.: G2dp: Diffusion planning with spatio-temporal grid guidance. arXiv preprint arXiv:2606.26017 (2026)
55. Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.S.: GA-Net: Guided aggregation net for end-to-end stereo matching. In: Image Processing On Line. pp. 185–194. IEEE (2019). https://doi.org/10.1109/CVPR.2019.00027
56. Zhu, L., Gao, Y., Zhang, J., Li, Y., Li, X.: Reliable and effective stereo matching for underwater scenes. Remote Sensing **16**(23), 4570 (2024). https://doi.org/10.3390/rs16234570