NBS: NO BIAS STEREO

Anonymous authors
Paper under double-blind review

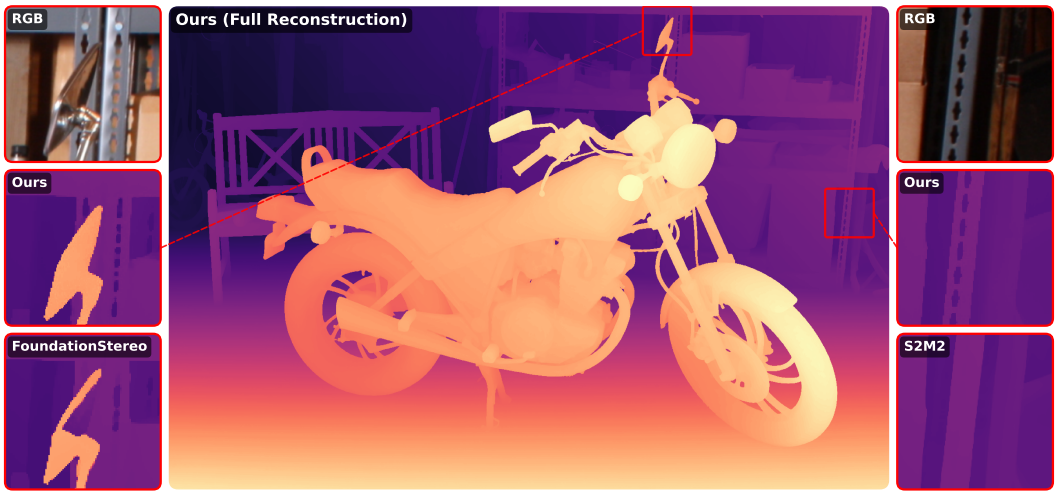


Figure 1: By removing complex inductive biases such as cost volumes and iterative GRU refinement, No Bias Stereo (NBS) accurately reconstructs fine details where prior methods fail. This result is generated using NBS-HighRes, evaluated zero-shot: it was never trained on Middlebury data, including this scene.

ABSTRACT

Stereo reconstruction is one of the last remaining computer vision tasks where state-of-the-art methods still rely on heavy, stereo-specific architectural inductive biases. Although it has been demonstrated that the task can be solved using general-purpose methods, such methods have since fallen behind specialized models, and it is widely believed that inductive biases in stereo are strictly necessary for both high-quality results and computational efficiency. We challenge this paradigm. In this paper, we demonstrate that both state-of-the-art accuracy and superior runtime efficiency are achievable with a model devoid of stereo-specific architectural inductive biases, relying instead on a simple, end-to-end Vision Transformer. By training on large-scale synthetic corpora, we show that a data-driven approach can surpass explicitly engineered geometry: our model ranks first on the ETH3D benchmark, leads on most metrics across four other benchmarks, and runs 4× faster than the previous best method. These results suggest that stereo-specific inductive biases are no longer a prerequisite for stereo matching, paving the way for scalable, general-purpose models for stereo reconstruction.

1 INTRODUCTION

Stereo reconstruction is a foundational computer vision task significantly advanced by deep learning over the last decade. During this period, the community has seen steady progress toward higher precision, driven by increasingly sophisticated architectural components. Modern state-of-the-art models (Wen et al., 2025b; Min et al., 2025a; Wang et al., 2024b; Min et al., 2025b; Li et al., 2022) rely on a highly specialized pipeline: deep feature extraction is followed by the construction of explicit 3D or 4D cost volumes, which are processed via 3D convolutions or filtered through iterative

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.

update blocks—most notably Gated Recurrent Units (GRUs)—to produce high-resolution disparity maps. As a result, each marginal gain has required increasingly complex, custom-engineered modules. Apart from ours, all non-anonymous methods in the top 30 of the ETH3D (Schöps et al., 2017) or Middlebury (Scharstein & Szeliski, 2002; Scharstein et al., 2014) benchmarks rely on strong architectural inductive biases. Consequently, the stereo vision community has somewhat diverged from the broader machine learning landscape, anchoring itself to highly specialized operations rather than embracing general-purpose architectures.

In this paper, we aim to bridge this gap. A central lesson of deep learning (Sutton, 2019) is that a simple, general-purpose architecture, combined with sufficient data and compute, ultimately surpasses highly specialized models. We show that stereo reconstruction is no exception. Concurrently, the field has recognized that integrating stereo matching with strong monocular depth priors can substantially reduce ambiguities in occluded or textureless regions. Researchers have begun integrating "internet-scale" knowledge by fusing pre-trained, often frozen monocular depth models into these pipelines, increasing their complexity (Wen et al., 2025b; Cheng et al., 2025; Jiang et al., 2025; Zhou et al., 2025). We propose that a single, end-to-end Vision Transformer (ViT) (Dosovitskiy et al., 2020) is a natural architecture to unify these two objectives. By eschewing rigid cost volumes and scaling on massive synthetic corpora, our transformer dynamically attends to both inter-image matching (stereo cues) and intra-image semantics (monocular priors). This pure attention mechanism naturally achieves the "best of both worlds" without specialized sub-modules or explicit geometric filtering.

Our model currently ranks first on the ETH3D benchmark and achieves state-of-the-art performance across several other benchmarks analyzed in this paper. Furthermore, by dispensing with computationally heavy cost volumes and iterative refinement steps, our architecture achieves highly efficient inference—operating several times faster than existing top-performing methods (measured at 966×546 px) while simultaneously delivering superior accuracy. A notable limitation, however, is scaling beyond our maximum training resolution of 1932×1330, as training at high resolutions remains computationally demanding for the standard self-attention mechanism.

Ultimately, we hope this work brings stereo reconstruction to the same paradigm as foundation models like Depth Anything 3 (DA3) (Lin et al., 2025) and VGGT (Wang et al., 2025a), motivating the community to move beyond custom inductive biases toward scalable, general-purpose architectures. To this end, our main contributions are as follows:

- **A Stereo Model Free of Stereo-Specific Inductive Biases:** We present a state-of-the-art stereo matching model devoid of explicit cost volumes, 3D convolutions, or iterative refinement, relying instead on a unified, end-to-end Vision Transformer architecture.
- **Emergent Monocular-Stereo Synthesis:** We demonstrate that our model internally learns to fuse geometric matching with monocular priors, achieving robust zero-shot generalization across diverse benchmarks and maintaining plausible depth estimation even when the right image is blanked out.
- **Extensive Evaluations:** We evaluate our approach on the standard Middlebury and ETH3D benchmarks. We also evaluate both our model and prior methods on additional high-fidelity synthetic data generated via SimpleProc (Ma et al., 2026), as well as the real-world XYZ-IBD (Huang et al., 2025) industrial part dataset post-processed for stereo. These evaluations demonstrate that data-driven learning can surpass explicitly engineered geometric biases. To support reproducibility, we attach our SimpleProc scene-generation seeds and post-processed, stereo-rectified XYZ-IBD pairs with ground-truth disparity in the supplement.

2 Related Work

2.1 Stereo

Classical stereo matching algorithms relied on local patch matching or semi-global optimization like SGM (Hirschmuller, 2008). Deep learning introduced a paradigm shift, replacing hand-crafted features with dense correlation volumes in architectures like GC-Net (Kendall et al., 2017) and DeepPruner (Duggal et al., 2019). To regularize these volumes, extensive research established dominant 3D CNN filtering architectures, including PSMNet (Chang & Chen, 2018), AANet (Xu & Zhang, 2020), CFNet (Shen et al., 2021), and PCW-Net (Shen et al., 2022). While accurate,

explicitly building and filtering 4D cost volumes is notoriously compute- and memory-intensive. To mitigate this bottleneck for real-time, high-resolution processing, lightweight alternatives like CoEx (Bangunharcana et al., 2021), fastACV (Xu et al., 2023b), and BGNet+ (Xu et al., 2021) were proposed, though often at the expense of high-resolution precision.

To bypass heavy 3D convolutions, RAFT-Stereo (Lipson et al., 2021) introduced a highly successful alternative: GRU-based iterative refinement over a 2D correlation field. This optical-flow-inspired approach was rapidly enhanced. CREStereo (Li et al., 2022) proposed cascaded refinement for real-world distortions, while hybrid architectures like IGEV-Stereo (Xu et al., 2023a) and IGEV++ (Xu et al., 2024) integrated geometry encoding volumes to guide disparity propagation. Subsequent works further refined this recurrent framework via dynamic frequency selection (Wang et al., 2024b), motif-based attention (Chen et al., 2024), and continuous disparity updates (Tosi et al., 2024). Despite these advances, the localized nature of iterative search limits the ability to resolve global ambiguities.

To overcome CNN locality bias, the field has increasingly integrated Vision Transformers (Dosovitskiy et al., 2020) and large-scale pretraining. However, early attention architectures like STTR (Li et al., 2021), CSTR (Guo et al., 2022), and UniMatch (Xu et al., 2023c) still rely on stereo-specific geometric modules like optimal transport layers and explicit correlation volumes. Meanwhile, structurally bias-free methods utilizing cross-view completion pretraining (Weinzaepfel et al., 2023; Kim et al., 2025) and full-context transformers (Bengana et al., 2022; Jia et al., 2024) have lagged significantly behind specialized models in accuracy and runtime. Consequently, to achieve SoTA performance and explicitly bridge the sim-to-real gap, recent frameworks (Wen et al., 2025b;a; Bartolomei et al., 2025; Min et al., 2025a; Wang & Deng, 2026) combine internet-scale monocular depth priors with explicit cost filtering and GRU refinement. In contrast, NBS is the first plain ViT, devoid of stereo-specific modules, to achieve both SoTA accuracy and superior efficiency.

2.2 Foundation Models

The broader vision community has embraced generalized foundation models that tackle diverse tasks using unified architectures. Initiated by the Vision Transformer (ViT) (Dosovitskiy et al., 2020), this trajectory extended to dense spatial prediction via DPT (Ranftl et al., 2021) and self-supervised models like DINOv2 (Oquab et al., 2024) and DINOv3 (Siméoni et al., 2025), which provide robust, task-agnostic features. Capitalizing on these representations, recent monocular geometry estimation frameworks (Yang et al., 2024; Hu et al., 2024; Bochkovskii et al., 2025; Wang et al., 2025b;c) have achieved high accuracy, laying the groundwork for unified multi-view geometric reasoning.

Leveraging these foundational representations, recent advancements focus on multi-view dense prediction. Feed-forward models like DUSt3R (Wang et al., 2024a) and its scalable successor Fast3R (Yang et al., 2025) abandon traditional explicit multi-view constraints for direct pointmap regression. Expanding on this paradigm, VGGT (Wang et al., 2025a) grounds transformer outputs in dense visual geometry, while DA3 (Lin et al., 2025) obviates architectural specialization, using a vanilla DINO transformer to predict unified depth-ray representations. Our approach aligns with this minimalist, data-driven trajectory. By replacing complex cost volume engineering with a standard ViT scaled on massive synthetic corpora, we show that simple feed-forward transformers are highly capable of stereo reconstruction. Moreover, we show that our approach estimates stereo depth with significantly higher precision than broader 3D foundation models.

3 Method

In this section, we discuss the model architecture illustrated in Figure 2, detailing our approach to high-resolution inference, the training data composition, and the overall training pipeline.

3.1 Task

We solve a standard stereo reconstruction task: given a pair of rectified images, we reconstruct a disparity map representing the horizontal pixel displacements between the left and right images. We assume disparity values are strictly positive and that disparities are well-defined even for regions in the left image that lack a direct match in the right image, as disparity is inherently a function of depth.

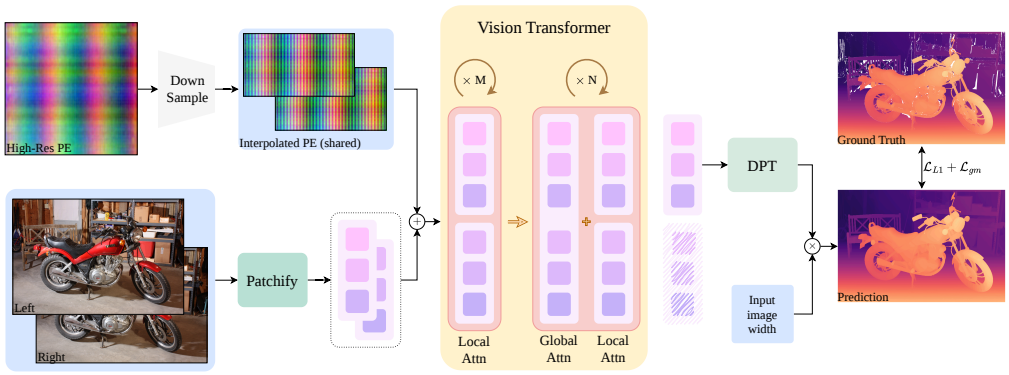


Figure 2: Our proposed architecture maintains a streamlined design. It leverages a standard Vision Transformer backbone – integrating local frame-wise and global self-attention mechanisms—alongside a DPT decoder. Previous state-of-the-art stereo methods use complex inductive biases such as frozen foundation models, multi-resolution cost volumes and iterative refinement via GRUs.

3.2 Model Architecture and High-Resolution Inference

Building on DINOv2 with registers (Darcet et al., 2024), we employ a standard Vision Transformer (ViT-Large) encoder, featuring local and global self-attention layers similar to VGGT and DA3. The combination of local and global self-attention allows the network to dynamically balance cross-image matching and intra-image semantics. This enables reliable disparity estimation across all parts of the image, including occluded regions or areas with no reliable match, such as transparent, reflective, or textureless surfaces. Specifically, our architecture employs local self-attention for layers 1 through 7, while every alternating layer from 8 to 24 is designated for global attention. For the initialization of the encoder weights, we leverage the original pretrained weights from DINOv2.

The task of disparity estimation requires very fine-grained positional understanding. To accommodate this, we find it highly beneficial to leverage positional embeddings with higher spatial resolution than the original DINOv2 model. Specifically, we use a 148×148 spatial resolution for 966×546 image resolution and 296×296 spatial resolution for 1932×1330 image resolution. Similar to DINOv2, the positional embeddings are resampled to the tokenized image's spatial resolution during the forward pass. We initialize the positional embeddings by upsampling the 37×37 DINOv2 pretrained embeddings.

To predict the disparity map, we utilize a Dense Prediction Transformer (DPT) decoder on the embeddings of the left input image. The DPT decoder effectively aggregates dense, multi-scale feature maps from the encoder to output fine-grained, edge-aware disparity predictions. The network is designed to output normalized disparity values between 0 and 1, which are subsequently multiplied by the image width to obtain the absolute disparity values in pixels. Forcing the model to generate normalized outputs relative to the image dimensions ensures that the learned disparity representations remain inherently independent of the specific input resolution.

The integration of a fine-tuning stage for specific, fixed higher resolutions (as explained in Section 3.4.2), together with the architectural choices described above, enables the model to seamlessly process images resized to the model's maximum native resolution of 1932×1330 during inference.

3.3 Internal Training Data

To support high-resolution scaling (Section 3.4.2), we augmented our high-resolution training mixture with a large-scale internal synthetic stereo dataset of approximately 2.4 million scenes. This dataset features stereo pairs with varying baselines, randomized camera intrinsics, diverse 3D assets, varied lighting conditions, material properties, and random object placements. Rendering was performed at three standard resolutions: 1024×768, 1280×720, and 1920×1080. Unlike the open-source stereo datasets utilized in our pipeline, this internal corpus introduces diversity in both resolution and

aspect ratio. Although this proprietary dataset will not be publicly released—and is therefore not claimed as a contribution of this work—we emphasize that it is used exclusively for high-resolution adaptation (Section 3.4.2); our base model (Section 3.4.1) and ETH3D submission rely solely on public datasets; the Middlebury, SimpleProc-M and XYZ-IBD results use models trained with the internal data (see Table 9 in Appendix for details).

3.4 Training Process

We employ a multi-stage training pipeline designed to progressively build strong geometric priors, adapt to high resolutions, and fine-tune for specific real-world benchmarks.

3.4.1 Base Model Pretraining

During the initial phase, we initialize the ViT-Large backbone with DINOv2 weights and pretrain our base model at a fixed resolution of 966×546, utilizing an upsampled 148×148 positional embedding grid. This stage is trained for 170,000 steps exclusively on the FSD (Wen et al., 2025b) dataset, distributing the workload across 128 A100 GPUs using mixed precision (bf16). We use the Muon-AdamW (Liu et al., 2025; Loshchilov & Hutter, 2019) optimizer with a peak learning rate of 1.46×10^{-4}, applying a 7,000-step linear warmup followed by a cosine annealing schedule.

3.4.2 High-Resolution Adaptation

To enable robust high-resolution inference, we branch from the base model in Section 3.4.1 and employ a multi-step pipeline that gradually scales the training resolution to a final target of 1932×1330. To accommodate the expanded token grid at this maximum resolution, the positional embeddings are progressively upsampled to a final spatial resolution of 296×296. Throughout this adaptation phase, the model is exposed to our massive, diverse mixture of 14 datasets, comprising our 2.4M scene internal dataset alongside 13 public datasets (Wen et al., 2025b; Tremblay et al., 2018; Bao et al., 2020; Jing et al., 2026; Mehl et al., 2023; Tosi et al., 2021; Li et al., 2022; Mayer et al., 2016; Cabon et al., 2020; Wang et al., 2020; Patel et al., 2025; Wang et al., 2021; Yan et al., 2025). To preserve the rich semantic features established during pretraining, we apply a differential learning rate strategy, scaling the ViT encoder's learning rate down to 10% of the overall rate. The intermediate scaling steps are trained using mixed precision (bf16) on 128 A100 GPUs, while the final high-resolution polish is conducted in full precision (float32) on 64 H100 GPUs to ensure peak predictive fidelity.

3.4.3 Benchmark-Specific Fine-Tuning

To achieve state-of-the-art results on specific real-world benchmarks, we perform a brief fine-tuning phase with a significantly reduced learning rate. For our ETH3D submission (NBS-ETH3D), we fine-tune the FSD-only base model (Section 3.4.1, 966×546) for 36,000 steps on FSD combined with the official training sets of ETH3D (Schöps et al., 2017) and Middlebury (Scharstein et al., 2014). For our Middlebury submission (NBS-Middlebury), we instead fine-tune NBS-HighRes (Section 3.4.2, 1932×1330) on the Middlebury training set plus its training mixture.

3.4.4 Training Objective

To optimize for global value accuracy, we employ a standard masked L1 loss between ground truth $\boldsymbol{d}$ and the prediction $\hat{\boldsymbol{d}}$, defined as:

$$\mathcal{L}_{\mathrm{L1}} = \frac{1}{\sum_i m_i} \sum_i m_i |\hat{d}_i - d_i|, \quad m_i = \mathbf{1}_{d_i > 0}. \tag{1}$$

To encourage locally consistent gradients while preserving sharp discontinuities in the prediction, we incorporate a multi-scale gradient matching term $\mathcal{L}_{\mathrm{gm}}$ (Li & Snavely, 2018; Ranftl et al., 2020). This is defined as an ℓ_1 penalty on the spatial gradients of the disparity difference matrix $\boldsymbol{R}^k = \hat{\boldsymbol{D}}^k - \boldsymbol{D}^k$ across scales k:

$$\mathcal{L}_{\mathrm{gm}} = \frac{1}{\sum_i m_i} \sum_i \sum_k m_i \left(|\nabla_x R_i^k| + |\nabla_y R_i^k| \right), \tag{2}$$

where R_i^k is the disparity difference map at position i and scale k. The final loss is the sum of these two terms. Appendix A.4 ablates the gradient-matching loss $\mathcal{L}_{\mathrm{gm}}$.

4 Results and Experiments

4.1 Benchmark Evaluation

We evaluate our models on the standard ETH3D and Middlebury benchmarks. For ETH3D, we resize images to our training resolution (966×546) and downsample the predicted disparities to the native scale. This resampling incurs no quantitative degradation. In contrast, on Middlebury, downsampling inputs to our maximum supported resolution (1932×1330) and upsampling predictions back to the original scale (up to 3000×2000) introduces a severe resolution bottleneck, which serves as the primary constraint on our Middlebury performance (see Section 5). Additionally, because the benchmarks contain few scenes, we extend our evaluation to two additional benchmarks designed to thoroughly test model generalizability. The first is a purely synthetic dataset generated via SimpleProc. Because none of the evaluated models (including ours) was trained on SimpleProc data, it serves as an "out-of-distribution" benchmark. Our second additional benchmark is derived from the XYZ-IBD dataset, repurposed specifically for stereo evaluation. To ensure full reproducibility, we include the processed XYZ-IBD dataset and the SimpleProc generation seeds in the supplement.

We use standard metrics in all evaluations: End-Point Error (EPE), Root Mean Square Error (RMS), and bad@$\{0.5, 1, 2, 4\}$ pixel error thresholds. We benchmark our models against CREStereo (Li et al., 2022), CroCo-Stereo (CroCo) (Weinzaepfel et al., 2023), Selective-IGEV (Sel-IGEV) (Wang et al., 2024b), FoundationStereo (FS) (Wen et al., 2025b), and S2M2-XL (S2M2) (Min et al., 2025a). Results are summarized in Table 1, and qualitative reconstructions are shown in Figure 3. Remaining metrics and model variants used for each evaluation are provided in the Appendix.

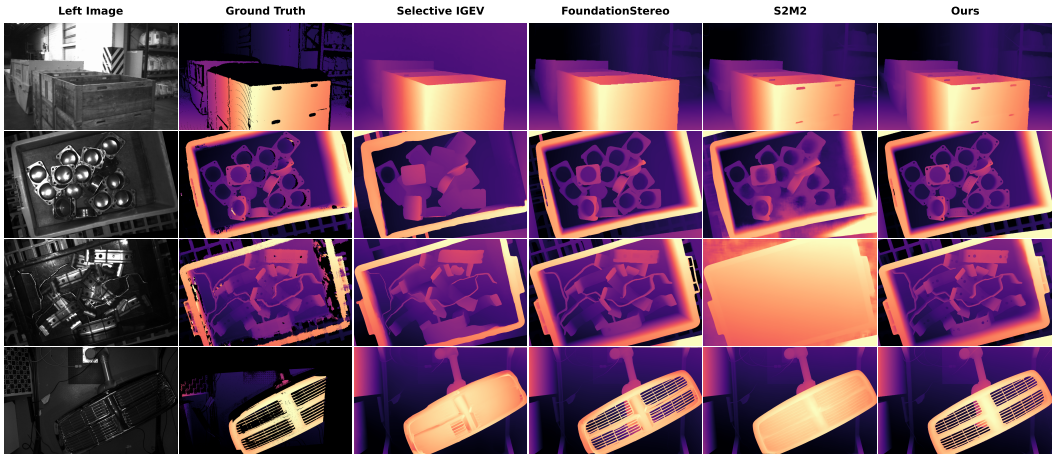


Figure 3: Our method achieves superior visual quality and state-of-the-art results, successfully reconstructing fine details that others miss, such as the holes in the containers (row 1), the complex shapes of cluttered objects (rows 2 and 3), and the shape and thin structures of the grill (row 4).

4.1.1 SimpleProc Evaluation Dataset

We utilized SimpleProc to generate a procedural dataset at two resolutions: SimpleProc-S with 661 scenes at 966×546 pixels, and SimpleProc-M with 530 scenes at 1932×1092 pixels. To tailor this tool for stereo at desired resolutions, we modified the default scene generation pipeline as explained in the Appendix. To prevent selection bias, all scenes were generated in a single batch without any manual filtering. While the randomized geometric primitives in these scenes do not test high-level semantic understanding, they provide a rigorous, unbiased assessment of a model's fundamental matching capabilities across a wide variety of unconstrained shapes and materials.

4.1.2 Industrial Parts Evaluation Dataset

Our second benchmark is constructed from XYZ-IBD, an open-source industrial parts dataset captured using a structured-light 3D scanner and a multi-camera setup. Although designed for 6-DoF pose

estimation, its high-fidelity captures of singulated and cluttered parts make it suited for stereo evaluation. From the five images per scene, we select and rectify the two best resembling a left-right stereo configuration. We then project all five structured-light scans onto the left view, fusing them into a ground-truth depth map via a per-pixel median operation. After excluding scenes with large artifacts, we establish a final evaluation set of 55 annotated stereo pairs, provided in the supplement.

4.1.3 Quantitative Results

As shown in Table 1, our model achieves SoTA performance on four of the five evaluation datasets, while remaining highly competitive on Middlebury despite the resolution downsampling. Whether evaluating on the real-world scenes of ETH3D, the challenging reflective surfaces of XYZ-IBD, or the procedurally generated geometries of SimpleProc, our inductive-bias-free architecture consistently exhibits superior matching precision. On SimpleProc, our model demonstrates remarkable robustness as the error margin widens; it outperforms all baselines. Notably, at the bad@4.0 metric, our model reduces the severe error rate of the next-best baseline, S2M2, by nearly 45%. This strong generalization across both real and synthetic domains validates the efficacy of our general-purpose attention mechanism over explicitly engineered geometry.

Table 1: Quantitative cross-dataset evaluation. We report EPE, bad@1.0, and bad@4.0 over all pixels across several benchmarks (RMS instead of bad@1.0 for Middlebury). Ours: NBS-ETH3D (ETH3D, test set), NBS-Middlebury (Middlebury, test set), NBS-Base (SimpleProc-S), NBS-HighRes (SimpleProc-M, XYZ-IBD).

Model	ETH3D			Middlebury			SimpleProc-S			SimpleProc-M			XYZ-IBD		
	EPE	b@1.0	b@4.0	EPE	RMS	b@4.0	EPE	b@1.0	b@4.0	EPE	b@1.0	b@4.0	EPE	b@1.0	b@4.0
CREStereo	0.14	1.09	0.12	2.10	10.50	5.05	0.33	4.72	1.43	0.64	8.03	2.75	15.92	62.28	36.04
CroCo	0.15	1.14	0.18	2.36	10.60	6.75	0.45	6.94	1.82	0.91	9.58	3.75	34.98	84.94	67.95
Sel-IGEV	0.15	1.56	0.28	1.54	9.26	3.74	0.52	6.13	2.07	0.75	8.16	3.19	17.48	60.92	34.98
FS	0.13	0.48	0.21	1.24	8.39	2.72	0.48	4.25	1.95	0.61	6.27	3.22	15.75	**54.69**	28.35
S2M2	0.10	0.26	0.04	**1.00**	7.39	**1.63**	0.29	4.16	1.12	0.44	6.04	2.31	19.24	57.22	30.69
Ours	**0.09**	**0.16**	**0.02**	1.05	**7.21**	1.73	**0.25**	**3.03**	**0.63**	**0.40**	**4.22**	**1.26**	**11.39**	55.61	**23.39**

Table 2 and Figure 4 present a comprehensive evaluation of model efficiency. All benchmarks were conducted on a single NVIDIA A100 GPU using float-32 precision, unless otherwise noted. Our method demonstrates superior performance, achieving SoTA results while substantially reducing latency and memory footprint. Furthermore, by leveraging Flash-Attention (Dao, 2023) and float-16 (F16) precision, we achieve a $5.4\times$ acceleration in inference speed and a $2\times$ reduction in memory usage, with only a marginal impact on accuracy.

Table 2: Efficiency on SimpleProc-S at 966×546 (one A100). Because our plain transformer maps efficiently onto modern GPU kernels, in F16 it is $4\times$ faster than S2M2 and $14.5\times$ faster than FS, with $2.9\times$ and $5.5\times$ lower peak memory, respectively.

Method	Params (M)	TFLOPs	Runtime (s)	Mem (GB)	bad@1
CREStereo	**9.5**	9.5	0.693	2.23	4.72
CroCo	436.2	75.5	3.778	3.88	6.94
Ours	351.5	**3.7**	0.323	2.52	**3.03**
Sel-IGEV (F16)	13.1	18.2	0.662	1.35	6.13
FS (F16)	374.5	25.9	0.872	6.74	4.25
S2M2 (F16)	405.7	11.3	0.240	3.52	4.16
Ours (F16)	351.5	**3.7**	**0.060**	**1.23**	3.24

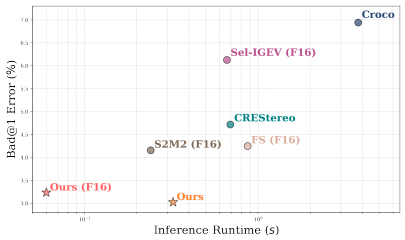


Figure 4: At 966×546 resolution, our simple architecture achieves superior runtime compared to previous SoTA models.

4.2 Emergent Monocular Depth Capabilities

We investigate the model's internal priors by evaluating its performance in a monocular depth estimation scenario. In this experiment, we provide a left image but replace the right image with an all-zero image, simulating a total right-camera occlusion. We benchmark our model against leading monocular depth estimation techniques (DA3, VGGT) and state-of-the-art stereo models (FS, S2M2) under identical conditions. To evaluate monocular depth capabilities, we first convert the predicted

disparity into depth. We then calculate the scale and shift via least-squares optimization, using the ground-truth depth as a reference. The resulting scaled depth is used to compute standard monocular metrics: Absolute Relative Error (AbsRel) and threshold accuracy δ_1. We evaluated our model on ETH3D, SimpleProc-S, and XYZ-IBD. As reported in Tables 3 and 4, our method achieves highly competitive monocular performance while maintaining SoTA stereo results. Notably, our approach narrows the gap between dedicated monocular estimators and stereo-based methods.

Our experiments reveal a trade-off in existing architectures: models trained primarily for monocular depth often underperform on stereo tasks, while leading stereo models typically struggle when restricted to monocular inputs (see Table 3). In contrast, our model preserves SoTA stereo precision while remaining highly competitive in monocular scenarios. As illustrated in Figure 5, the model produces physically plausible disparity maps even under total right-camera occlusion, suggesting it has developed a robust internal representation of geometric priors.

Table 3: Our monocular depth estimation rivals DA3 and VGGT while outperforming stereo method S2M2. FS uses a frozen DAv2 (Yang et al., 2024) trained for mono-depth.

Model	ETH3D		SimpleProc-S		XYZ-IBD	
	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
DA3Large	0.108	0.896	0.708	0.614	3.413	0.137
DA3Metric	*0.068*	*0.946*	0.699	0.641	3.383	0.127
VGGT	**0.065**	**0.956**	0.694	**0.685**	3.624	0.140
S2M2	0.220	0.666	0.824	0.569	3.389	*0.144*
FS (DAv2)	0.160	0.790	*0.586*	*0.643*	**2.700**	0.140
Ours	0.164	0.773	**0.470**	0.628	*2.788*	**0.172**

Table 4: When evaluating stereo depth estimation metrics instead of disparity estimation, our model still achieves state-of-the-art results across most metrics and benchmarks.

Model	ETH3D		SimpleProc-S		XYZ-IBD	
	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
DA3Large	2269.640	0.000	3357.060	0.000	991.903	0.000
DA3Metric	0.238	0.814	4.640	0.175	1.601	0.035
VGGT	0.299	0.967	79.935	0.001	0.862	0.000
S2M2	0.035	*0.995*	**0.026**	*0.987*	0.203	0.813
FS	*0.028*	*0.995*	0.059	0.977	*0.132*	*0.830*
Ours	**0.021**	**0.998**	*0.030*	**0.989**	**0.128**	**0.882**

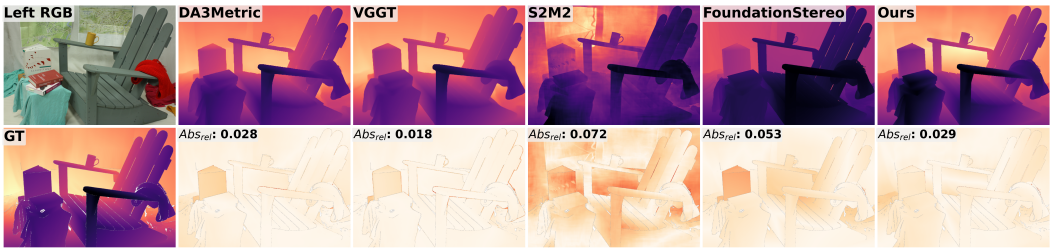


Figure 5: Our model (Col. 6) demonstrates superior visual quality for monocular depth when compared to SoTA stereo methods S2M2 (Col. 4) and FS (Col. 5) while being competitive with SoTA monodepth models DA3Metric (Col. 2) and VGGT (Col. 3).

4.3 Comparison with 3D Foundation Models

Architecturally, our approach shares similarities with recent 3D foundation models, including Depth Anything 3 and VGGT. However, the primary distinction lies in our output representation. To provide a fair comparison across these different paradigms, in Table 4 we evaluate all models on depth-based metrics by converting our predicted disparity maps into absolute depth using camera parameters. The results highlight our model's competitive performance in both zero-shot generalization and high-frequency detail preservation compared to these broad-purpose foundation models.

4.4 Positional Embedding Spatial Resolution Analysis

We evaluate the impact of different positional embedding spatial resolutions on disparity estimation. For this ablation, we trained a model exclusively on the FSD dataset using two spatial resolutions: 37×37 and 148×148. Performance was measured on a random 1% validation split from FSD. Quantitatively, the high-capacity PE grid yields robust performance gains. When inferring at a 966×546 scale, the 148×148 grid reduced the strict bad@0.5 error by an absolute 7.21 points compared to the 37×37 baseline. Similar error reductions were observed at the 518×280 scale (bad@0.5 reduced by 3.83 points), confirming that high-resolution spatial priors are robust and

universally beneficial across varying inference dimensions. Qualitatively (Figure 6), the 37×37 embeddings successfully capture low-frequency spatial priors, but are insufficient to accurately estimate sub-pixel disparities on a 966 pixel grid. The 148×148 positional embeddings allow the network to learn refined, high-frequency spatial variations. This density provides the "spatial real estate" necessary to encode localized coordinate systems and resolve fine structures.

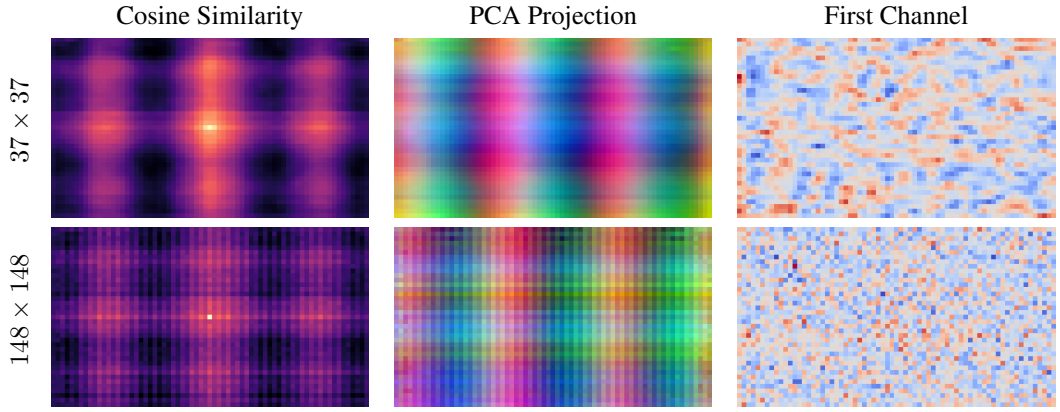


Figure 6: As the model trains, the learned 148×148 positional embeddings (Bottom) develop much sharper spatial structure than the 37×37 embeddings (Top), which in turn allows significantly better disparity estimates. We visualize the cosine similarity with the center pixel (Left), the PCA projection (Middle), and the weight of the first channel (Right).

5 Limitations

While our inductive-bias-free architecture demonstrates strong performance, it currently exhibits a few limitations. First, the model is trained strictly to predict positive disparities; consequently, it cannot resolve negative disparities that occasionally occur in real-world deployments. Second, because our final model weights are fine-tuned for specific, fixed spatial resolutions, input images must be resized or padded to these exact dimensions before inference. This requirement restricts the model's flexibility compared to architectures that dynamically process arbitrary image sizes and aspect ratios on the fly. Finally, similar to DA3 and VGGT, our reliance on a standard ViT backbone introduces a quadratic memory bottleneck during training at extreme resolutions. Because our architecture must be fine-tuned at a specific spatial dimension for native inference, we cannot currently evaluate ultra-high-resolution datasets—such as Middlebury (up to 3000×2000) and Booster (Ramirez et al., 2024) (4112×3008)—at their native scales. Fully convolutional or iterative networks, though highly specialized, can process these extreme dimensions during inference without retraining, whereas our current setup requires downsampling them to our maximum trained resolution (1932×1330). Scaling our native training pipeline to accommodate these extreme dimensions remains an important area for future work. Nevertheless, by relying entirely on general-purpose attention rather than custom geometry modules, our model is well positioned to benefit from the rapidly expanding ViT optimization ecosystem, including token merging, efficient attention, and quantization.

6 Conclusion

In this work, we challenged the long-standing paradigm that explicit architectural inductive biases—such as 3D convolutions, cost volumes, and iterative refinement—are strictly necessary for state-of-the-art stereo reconstruction. By scaling a simple, end-to-end Vision Transformer on a massive synthetic dataset, we demonstrated that pure data-driven learning can match and surpass explicitly engineered geometry. Our model achieves highly competitive accuracy across standard benchmarks and naturally learns to fuse inter-image matching with intra-image monocular priors without requiring specialized sub-modules. We hope this work serves as a stepping stone, encouraging the community to move beyond highly specialized stereo pipelines and stimulating further research into unified, general-purpose architectures for 3D reconstruction.

AI USE STATEMENT

In this work, we have not used generative AI tools for generating synthetic datasets, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results.

Furthermore, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in writing proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, and assisting with translation are not applicable to this work.

Additionally, we utilized generative AI tools to assist in creating and editing software code, as well as editing the manuscript to improve readability. We have thoroughly reviewed all AI-assisted work; specifically, all text went through multiple stages of human proofreading and polishing, and all generated code passed human code review. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we provide comprehensive details of our model architecture, datasets, and experimental protocols. The complete model architecture and training pipeline are described in Section 3, with full training and inference compute specifications detailed in Section 3.4 and Table 2, respectively. All internal experimental settings and hyperparameters are documented in Section 3.4. To facilitate replication of our evaluations, we attach the post-processed XYZ-IBD industrial parts dataset along with the exact procedural generation seeds for SimpleProc in the supplement. We also provide scene-generation instructions for SimpleProc in the Appendix. For public benchmarks (ETH3D and Middlebury), we adhere strictly to their standardized, blind submission protocols and refer to the original benchmark publications for dataset details.

REFERENCES

Antyanta Bangunharcana, Jae Won Cho, Seokju Lee, In So Kweon, Kyung-Soo Kim, and Soohyun Kim. Correlate-and-excite: Real-time stereo matching via guided cost volume excitation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3542–3548. IEEE, 2021.

Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63(11):212101, 2020.

Luca Bartolomei, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1013–1027, 2025.

Nadir Bengana, Janne Mustaniemi, and Janne Heikkilä. Seeking attention: Using full context transformers for better disparity estimation. In *International Conference on Pattern Recognition and Artificial Intelligence*, pp. 398–409. Springer, 2022.

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations (ICLR)*, 2025.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. URL `https://arxiv.org/abs/2001.10773`.

Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5410–5418, 2018.

Ziyang Chen, Wei Long, He Yao, Yongjun Zhang, Bingshu Wang, Yongbin Qin, and Jia Wu. Mocha-stereo: Motif channel attention network for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27768–27777, 2024.

Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, and Xin Yang. Monster: Marry monodepth to stereo unleashes power. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6273–6282, 2025.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4384–4393, 2019.

Weiyu Guo, Zhaoshuo Li, Yongkui Yang, Zheng Wang, Russell H Taylor, Mathias Unberath, Alan Yuille, and Yingwei Li. Context-enhanced stereo transformer. In *European Conference on Computer Vision*, pp. 263–279. Springer, 2022.

Heiko Hirschmuller. Stereo processing by semi-global matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2008.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Kaixuan Wang, Hao Chen, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3dv2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Junwen Huang, Jiaqi Hu, Peter KT Yu, Slobodan Ilic, Martin Sundermeyer, and Benjamin Busam. Xyz-ibd: High-precision bin-picking dataset for object 6d pose estimation capturing real-world industrial complexity, 2025. URL `https://arxiv.org/abs/2506.00599`.

Di Jia, Peng Cai, Qian Wang, and Ninghua Yang. A transformer-based architecture for high-resolution stereo matching. *IEEE Transactions on Computational Imaging*, 10:83–92, 2024.

Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. Defom-stereo: Depth foundation model based stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21857–21867, 2025.

Junpeng Jing, Ye Mao, Anlan Qiu, and Krystian Mikolajczyk. Match stereo videos via bidirectional alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pp. 66–75, 2017.

Soomin Kim, Hyesong Choi, Jihye Ahn, and Dongbo Min. Unitt-stereo: Unified training of transformer for enhanced stereo matching. *IEEE Access*, 13:204695–204707, 2025.

Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16263–16272, 2022.

Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6197–6206, 2021.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2041–2050, 2018.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*, pp. 218–227. IEEE, 2021.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training, 2025. URL https://arxiv.org/abs/2502.16982.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Zeyu Ma, Alexander Raistrick, and Jia Deng. Simpleproc: Fully procedural synthetic data from simple rules for multi-view stereo, 2026. URL https://arxiv.org/abs/2604.04925.

N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL http://lmb.informatik.uni-freiburg.de/Publications/2016/MIFDB16. arXiv:1512.02134.

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4981–4991, 2023.

Junhong Min, Youngpil Jeon, Jimin Kim, and Minyong Choi. S2m2: Scalable stereo matching model for reliable depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26729–26739, 2025a.

Junhong Min, Jimin Kim, Minwook Kim, Cheol-Hui Min, Youngpil Jeon, and Minyong Choi. Depthfocus: Controllable depth estimation for see-through scenes. *arXiv preprint arXiv:2511.16993*, 2025b.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian Scherer, Marco Hutter, and Wenshan Wang. Tartanground: A large-scale dataset for ground robot perception and navigation. *arXiv preprint arXiv:2505.10696*, 2025.

Pierluigi Zama Ramirez, Alex Costanzino, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Booster: A benchmark for depth from images of specular and transparent surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):85–102, 2024. doi: 10.1109/TPAMI.2023.3323858.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 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, 2020.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168. IEEE, 2021.

Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002.

Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pp. 31–42. Springer, 2014.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13906–13915, 2021.

Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *European Conference on Computer Vision*, pp. 280–297. Springer, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Richard Sutton. The bitter lesson. `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`, 2019. Incomplete Ideas (blog).

Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8942–8952, 2021.

Fabio Tosi, Filippo Aleotti, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Neural disparity refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation, 2018. URL `https://arxiv.org/abs/1804.06534`.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025a.

Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2021. doi: 10.1109/ICME51207.2021.9428423.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5261–5271, 2025b.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 35928–35959. Curran Associates, Inc., 2025c. doi: 10.52202/085713-1207. URL `https://proceedings.neurips.cc/paper_files/paper/2025/file/336572db3e99930814d6b328d4220cb6-Paper-Conference.pdf`.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024a.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE Press, 2020. doi: 10.1109/IROS45743.2020.9341801. URL https://doi.org/10.1109/IROS45743.2020.9341801.

Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19701–19710, 2024b.

Yihan Wang and Jia Deng. Waft-stereo: Warping-alone field transforms for stereo matching. *arXiv preprint arXiv:2603.24836*, 2026.

Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17969–17980, 2023.

Bowen Wen, Shaurya Dewan, and Stan Birchfield. Fast-foundationstereo: Real-time zero-shot stereo matching. *arXiv preprint arXiv:2512.11130*, 2025a.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5249–5260, 2025b.

Bin Xu, Yuhua Xu, Xiaoli Yang, Wei Jia, and Yulan Guo. Bilateral grid learning for stereo matching networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12497–12506, 2021.

Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21919–21928, 2023a.

Gangwei Xu, Yun Wang, Junda Cheng, Jinhui Tang, and Xin Yang. Accurate and efficient stereo matching via attention concatenation volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2461–2474, 2023b.

Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Junda Cheng, Chunyuan Liao, and Xin Yang. Igev++: Iterative multi-range geometry encoding volumes for stereo matching. *arXiv preprint arXiv:2409.00638*, 2024.

Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1959–1968, 2020.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023c.

David Yan, Alexander Raistrick, and Jia Deng. What makes good synthetic training data for zero-shot stereo matching?, 2025. URL https://arxiv.org/abs/2504.16930.

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, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.

Jingyi Zhou, Haoyu Zhang, Jiakang Yuan, Peng Ye, Tao Chen, Hao Jiang, Meiya Chen, and Yangyang Zhang. All-in-one: Transferring vision foundation models into stereo matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10797–10805, 2025.

A Appendix

A.1 Full Evaluation Results

In this section, we present comprehensive quantitative evaluations across four benchmarks: XYZ-IBD (Huang et al., 2025), SimpleProc-S (Ma et al., 2026), SimpleProc-M (Ma et al., 2026), and ETH3D (Schöps et al., 2017). Due to space constraints in the main text, the full tables are detailed here to provide a complete picture of our method's performance relative to recent state-of-the-art architectures.

Across all four benchmarks, our proposed method demonstrates state-of-the-art or highly competitive performance, exhibiting strong generalization capabilities across synthetic, real-world, indoor, and outdoor settings:

- **XYZ-IBD** (Table 5): On this challenging dataset, our model achieves top-tier results. It secures the absolute best performance in End-Point Error (EPE), bad@2 (**36.64**) and bad@4 (**23.39**), while remaining highly competitive across all other error thresholds.
- **SimpleProc-S / SimpleProc-M** (Table 6): Evaluating across different image resolutions demonstrates our method's multi-scale robustness. At the 966×546 resolution (SimpleProc-S), our model ranks first on every metric except bad@0.5, where it is second to FoundationStereo. At the higher 1932×1092 resolution (SimpleProc-M), it maintains dominance, achieving the lowest EPE (**0.40**) and reducing the bad@4 error by $\approx 45\%$ relative to the second-best model (1.26 vs 2.31).
- **ETH3D** (Table 7): Our framework establishes a new state-of-the-art across virtually all metrics on the ETH3D test set. Notably, for non-occluded pixels, it achieves an EPE of **0.09** and slashes the bad@0.5 metric down to **0.65** (nearly half the error rate of FoundationStereo (Wen et al., 2025b)). This exceptional sub-pixel accuracy is consistently maintained when evaluating all pixels as well.

Table 5: Quantitative comparison on the XYZ-IBD dataset (Huang et al., 2025). The second-best values are highlighted with a superscript 2.

Model	EPE	RMS	bad@0.5	bad@1	bad@2	bad@4
CroCo	34.98	47.77	91.59	84.94	76.69	67.95
CREStereo	15.92	**34.36**	77.65	62.28	46.96	36.04
FS	15.75[2]	35.87	**72.63**	**54.69**	38.43[2]	28.35[2]
Sel-IGEV	17.48	35.44[2]	76.98	60.92	45.53	34.98
S2M2	19.24	49.05	75.26	57.22	40.74	30.69
Ours	**11.39**	38.21	74.22[2]	55.61[2]	**36.64**	**23.39**

Table 6: Quantitative comparison on SimpleProc (Ma et al., 2026) dataset at different resolutions on all pixels. The second-best values are highlighted with a superscript 2.

Model	SimpleProc-S						SimpleProc-M					
	EPE	RMS	bad@0.5	bad@1.0	bad@2.0	bad@4.0	EPE	RMS	bad@0.5	bad@1.0	bad@2.0	bad@4.0
CREStereo	0.33	1.30	8.32	4.72	2.58	1.43	0.64	2.70	12.87	8.03	4.89	2.75
CroCo	0.45	1.43	14.69	6.94	3.50	1.82	0.91	3.11	17.14	9.58	5.83	3.75
FS	0.48	2.50	**7.23**	4.25	2.79	1.95	0.61	2.82	9.49[2]	6.27	4.47	3.22
Sel-IGEV	0.52	2.15	11.73	6.13	3.46	2.07	0.75	3.36	13.49	8.16	5.13	3.19
S2M2	0.29[2]	1.00[2]	7.97	4.16[2]	2.20[2]	1.12[2]	0.44[2]	**1.73**	9.69	6.04[2]	3.83[2]	2.31[2]
Ours	**0.25**	**0.80**	7.83[2]	**3.03**	**1.36**	**0.63**	**0.40**	1.91[2]	**8.59**	**4.22**	**2.26**	**1.26**

Table 7: Results on the test set of ETH3D. Our method achieves state-of-the-art performance across both non-occluded and all-pixel regions.

Model	Non-occluded pixels						All pixels					
	EPE	RMS	bad@0.5	bad@1	bad@2	bad@4	EPE	RMS	bad@0.5	bad@1	bad@2	bad@4
RAFT-Stereo	0.18	0.36	7.04	2.44	0.44	0.15	0.19	0.42	7.33	2.60	0.56	0.22
CroCo	0.14	0.30	3.27	0.99	0.39	0.13	0.15	0.35	3.51	1.14	0.50	0.18
CREStereo	0.13	0.28	3.58	0.98	0.22	0.10	0.14	0.31	3.75	1.09	0.29	0.12
FS	**0.09**	0.20	1.26	0.26	0.08	0.05	0.13	0.61	1.56	0.48	0.26	0.21
S2M2	0.10	0.20	0.93	0.22	0.06	0.03	0.10	0.22	1.03	0.26	0.08	0.04
WAFT-Stereo	0.10	0.20	0.89	0.28	0.07	0.03	0.10	0.21	0.97	0.31	0.07	0.03
DepthFocus	0.10	0.19	1.04	0.25	0.07	**0.02**	0.10	0.22	1.14	0.29	0.09	0.04
Ours	**0.09**	**0.17**	**0.65**	**0.14**	**0.04**	**0.02**	**0.09**	**0.18**	**0.71**	**0.16**	**0.04**	**0.02**

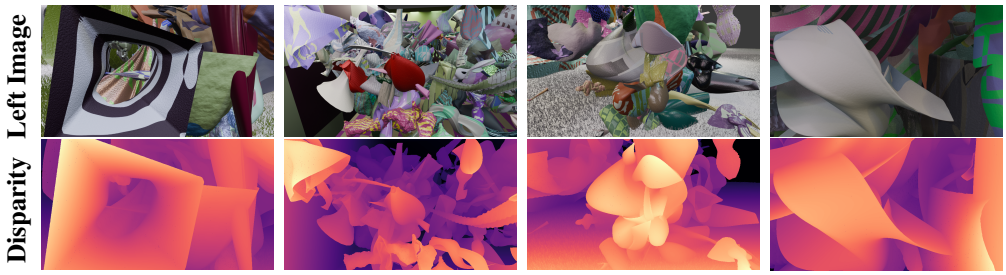


Figure 7: Visualizing multiple scenes from 966×546 SimpleProc evaluation dataset.

A.2 Generating SimpleProc Evaluation Dataset

The following modifications are required from the original SimpleProc data generation to reproduce the evaluation dataset:

- Modify *fov_y* and *total_pixels* parameters according to the desired width and height of the render. Disable rounding of the image resolution.
- Place an anchor rig (which forms the right camera of the stereo pair) using the default pipeline and parameter sampling.
- For the left rectified camera rig, we compute its location and rotation using the anchor rig and a uniformly sampled baseline shift (in the range 5–45 cm) as:
  $right_vec \leftarrow anchor_rig.matrix_world_{3\times3} \cdot [1, 0, 0]^T$
  $cam_location \leftarrow anchor_rig.location - (baseline \cdot normalized_right_vec)$
  $rig.rotation_euler \leftarrow anchor_rig.rotation_euler$
- For both camera rigs, the parent inverse of the camera object is set to identity so local camera coordinates map $1:1$ to the rig.
- Each scene render is performed with a unique seed value. The list of seed values used for each resolution is attached in the supplement.

Example images of this generated dataset can be found in Figure 7.

A.3 Industrial Parts Evaluation Dataset

We include the industrial parts evaluation dataset, which is XYZ-IBD (Huang et al., 2025) post-processed for stereo, as shown in Figure 8. We include an *xyz.zip* file containing 55 folders in the supplement. Each folder contains *left_image.jpeg*, *right_image.jpeg*, *disp.png*, and *calib.json*. The left and right images are already rectified and resized to 1080p. The ground truth disparity is saved in uint16 png. To get the true disparity, divide all values by 2. The calibration contains the output of cv2.stereoRectify and the baseline, which can be used to construct a 3D pointcloud.
Although the ground truth for these scenes occasionally exhibits artifacts from the structured-light cameras employed, we find this negligible, since these inconsistencies impact all evaluated methods in the same manner.

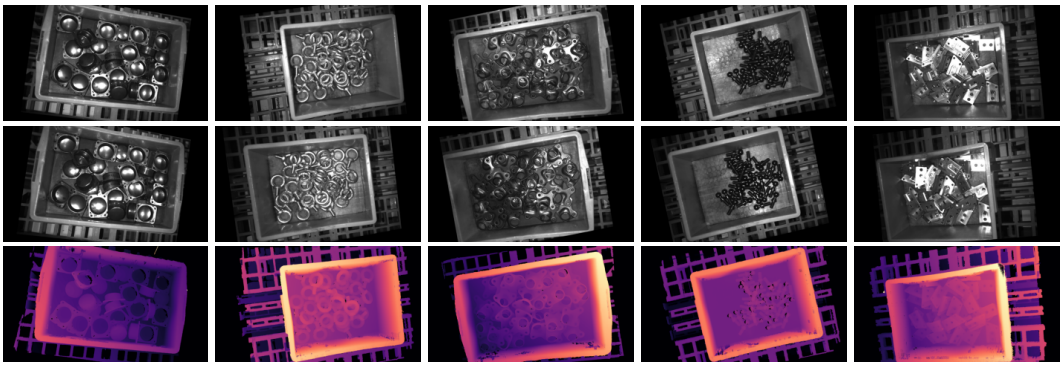


Figure 8: Examples from XYZ-IBD (Huang et al., 2025) dataset. Rows show left images, right images, and ground-truth disparity.

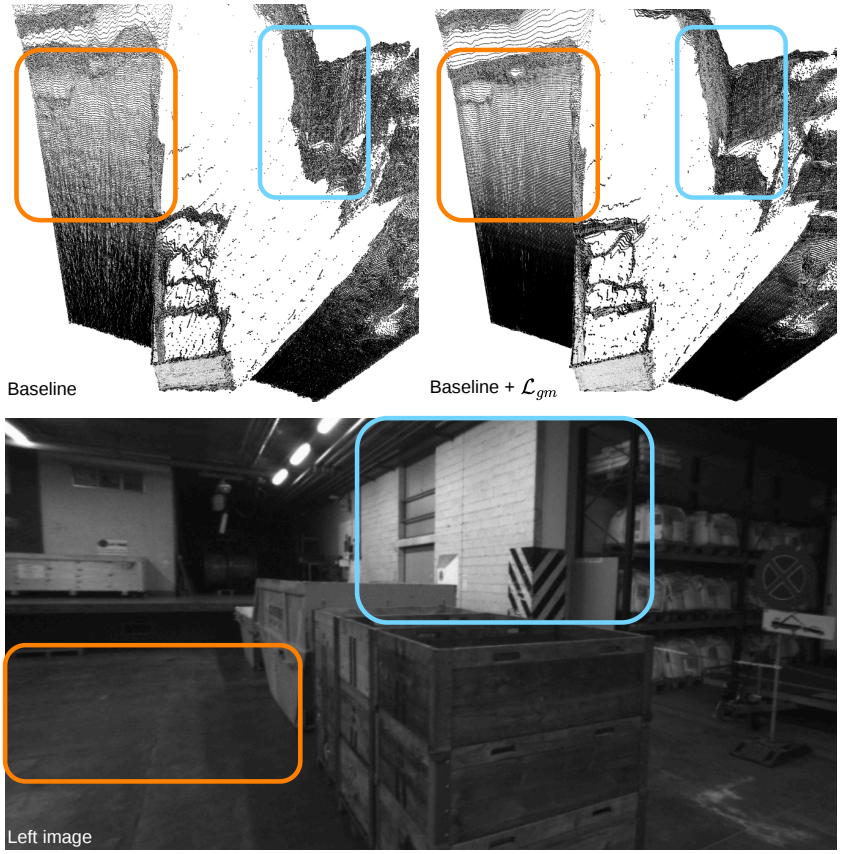


Figure 9: Qualitative comparison of disparity predictions with and without the Gradient Matching loss. The point clouds are visualized in the first row and the left image is provided in the second row.

A.4 Ablation on Gradient Matching Loss

Due to the significant computational requirements of training our full-scale model, this experiment was performed using a smaller backbone (ViT-Base) and a data subset. To validate the effectiveness of the gradient matching loss, we trained this model exclusively on the FSD dataset and evaluated it on the ETH3D training split. Table 8 and Figure 9 summarize the ablation results. Although the impact on global metrics is minor, the qualitative results in Figure 9 demonstrate that $\mathcal{L}_{gm}$ is essential for suppressing noise on the geometry reconstructions. With $\mathcal{L}_{gm}$, flat surfaces such as the ground and walls are less noisy, and depth discontinuities stay sharp instead of blurred.

Table 8: Ablation study on the Gradient Matching loss ($\mathcal{L}_{gm}$). While global metrics remain stable, $\mathcal{L}_{gm}$ significantly enhances local structural sharpness.

Method	EPE ↓	RMS ↓	bad@0.5 ↓
ViT-Base (Baseline)	**0.282**	0.545	13.56
ViT-Base + $\mathcal{L}_{gm}$	0.285	**0.525**	**13.51**

Table 9: Comparison of total training stereo pairs across models.

Model	Total Stereo Pairs	Target Evaluations
FoundationStereo	1.43M	All benchmarks
S2M2	2M+	All benchmarks
NBS-Base (966×546)	1.11M	SimpleProc-S, Monodepth
NBS-ETH3D (966×546)	1.11M + 300 (Middlebury) + 27 (ETH3D)	ETH3D
NBS-HighRes (1932×1330)	5.87M	SimpleProc-M, XYZ-IBD
NBS-Middlebury (1932×1330)	5.87M + 300 (Middlebury)	Middlebury

A.5 Training Data Volume Comparison

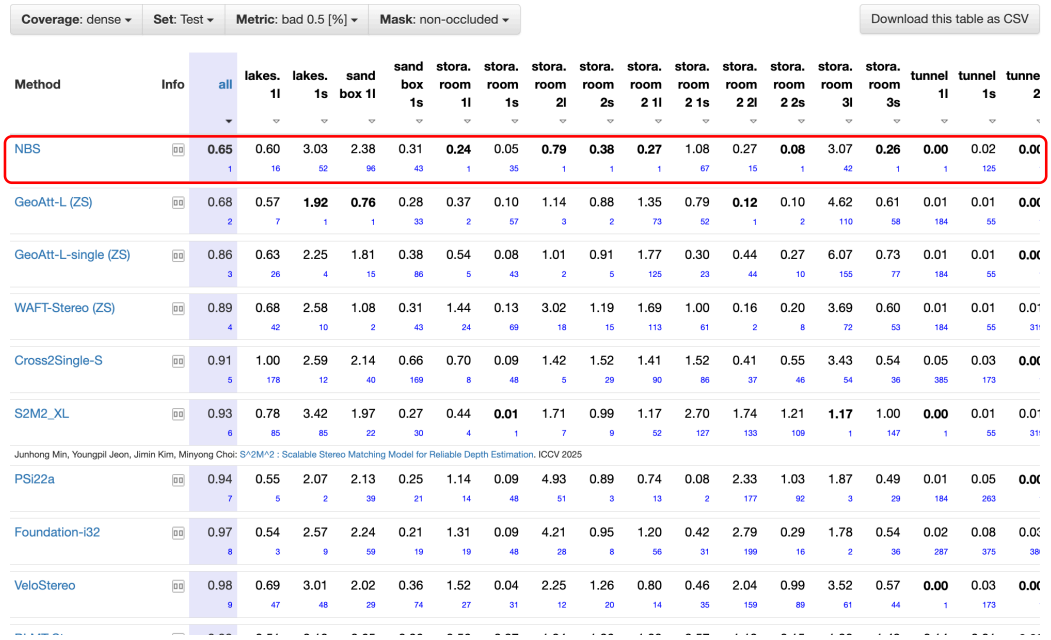
To explicitly disentangle the effects of our bias-free Vision Transformer architecture from training data scale, Table 9 compares the exact training data composition across our NBS model variants and current state-of-the-art methods, FoundationStereo (Wen et al., 2025b) and S2M2 (Min et al., 2025a).

Crucially, both NBS-Base and NBS-ETH3D utilize 22% less training data than FoundationStereo and 45% less than S2M2, while achieving the SoTA performance on the corresponding benchmarks. The expanded mixture (3.47M public pairs + 2.40M internal pairs) is strictly restricted to the high-resolution adaptation stage (NBS-HighRes, 1932×1330).

In Figure 1, we present results using NBS-HighRes rather than NBS-Middlebury. This choice is motivated by structural artifacts present in the Middlebury ground-truth disparities – specifically, the erroneous filling of fine cavities and thin structures – whereas NBS-HighRes provides significantly higher fidelity around delicate geometric boundaries.

A.6 ETH3D Leaderboard

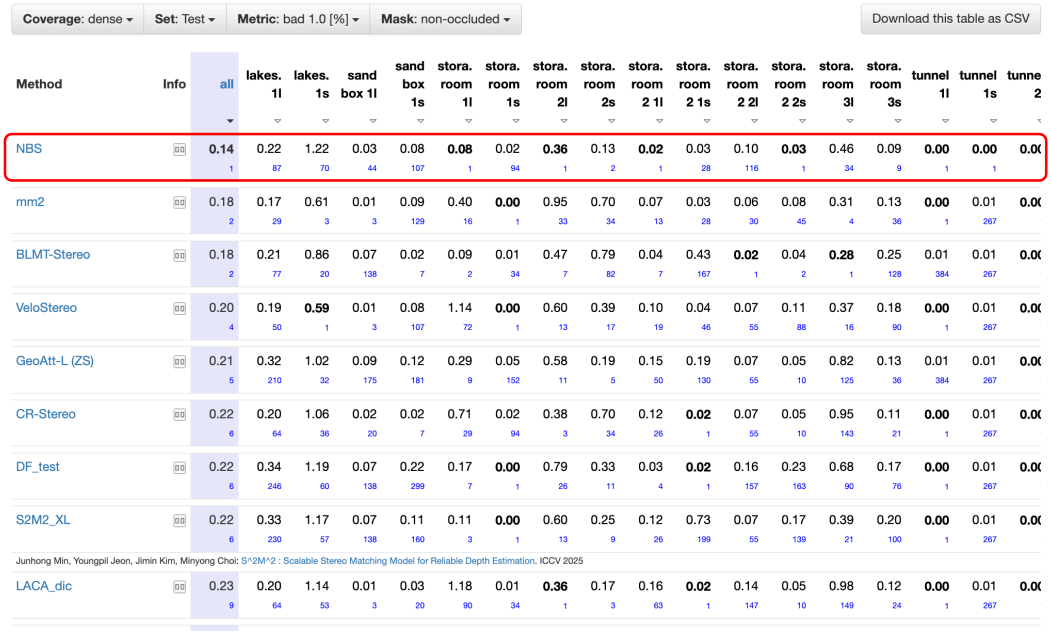
At the time of submission, our method achieved first place on the ETH3D Leaderboard. Figures 10, 11, 12, and 13 demonstrate that our model consistently outperforms existing approaches across multiple metrics by a substantial margin. We provide screenshots of only a subset of the metrics, while Table 7 shows the results for all metrics.



Coverage: dense | Set: Test | Metric: bad 0.5 [%] | Mask: non-occluded | Download this table as CSV

Method	Info	all	lakes. 1l	lakes. 1s	sand box 1l	sand box 1s	stora. room 1l	stora. room 1s	stora. room 2l	stora. room 2s	stora. room 2 1l	stora. room 2 1s	stora. room 2 2l	stora. room 2 2s	stora. room 3l	stora. room 3s	tunnel 1l	tunnel 1s	tunne 2
NBS		**0.65** 1	0.60 16	3.03 52	2.38 96	0.31 43	**0.24** 1	0.05 35	**0.79** 1	**0.38** 1	**0.27** 1	1.08 67	0.27 15	**0.08** 1	3.07 42	**0.26** 1	**0.00** 1	0.02 125	**0.0**
GeoAtt-L (ZS)		0.68 2	0.57 7	**1.92** 1	**0.76** 1	0.28 33	0.37 2	0.10 57	1.14 3	0.88 2	1.35 73	0.79 52	**0.12** 1	0.10 2	4.62 110	0.61 58	0.01 184	0.01 55	**0.0**
GeoAtt-L-single (ZS)		0.86 3	0.63 26	2.25 4	1.81 15	0.38 86	0.54 5	0.08 43	1.01 2	0.91 5	1.77 125	0.30 23	0.44 44	0.27 10	6.07 155	0.73 77	0.01 184	0.01 55	**0.0**
WAFT-Stereo (ZS)		0.89 4	0.68 42	2.58 10	1.08 2	0.31 43	1.44 24	0.13 69	3.02 18	1.19 15	1.69 113	1.00 61	0.16 2	0.20 8	3.69 72	0.60 53	0.01 184	0.01 55	0.0 31
Cross2Single-S		0.91 5	1.00 178	2.59 12	2.14 40	0.66 169	0.70 8	0.09 48	1.42 5	1.52 29	1.41 90	1.52 86	0.41 37	0.55 46	3.43 54	0.54 36	0.05 385	0.03 173	**0.0**
S2M2_XL		0.93 6	0.78 85	3.42 85	1.97 22	0.27 30	0.44 4	**0.01** 1	1.71 7	0.99 9	1.17 52	2.70 127	1.74 133	1.21 109	**1.17** 1	1.00 147	**0.00** 1	0.01 55	0.0 31
Junhong Min, Youngpil Jeon, Jimin Kim, Minyong Choi: S^2M^2 : Scalable Stereo Matching Model for Reliable Depth Estimation. ICCV 2025																			
PSi22a		0.94 7	0.55 5	2.07 2	2.13 39	0.25 21	1.14 14	0.09 48	4.93 51	0.89 3	0.74 13	0.08 2	2.33 177	1.03 92	1.87 3	0.49 29	0.01 184	0.05 263	**0.0**
Foundation-i32		0.97 8	0.54 3	2.57 9	2.24 59	0.21 19	1.31 19	0.09 48	4.21 28	0.95 8	1.20 56	0.42 31	2.79 199	0.29 16	1.78 2	0.54 36	0.02 287	0.08 375	0.0 38
VeloStereo		0.98 9	0.69 47	3.01 48	2.02 29	0.36 74	1.52 27	0.04 31	2.25 12	1.26 20	0.80 14	0.46 35	2.04 159	0.99 89	3.52 61	0.57 44	**0.00** 1	0.03 173	**0.0**

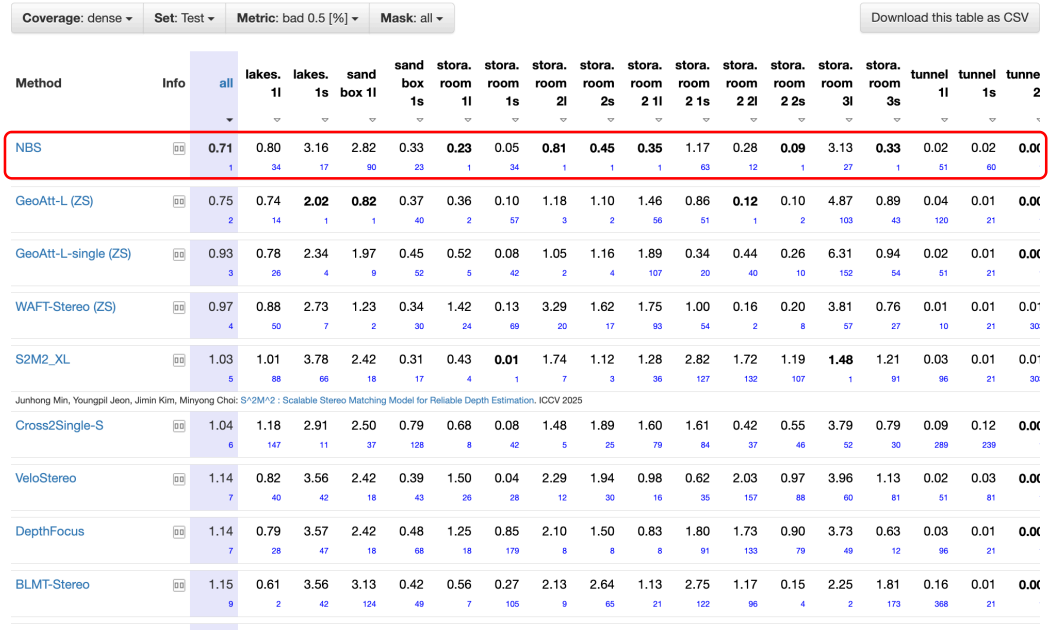
Figure 10: ETH3D leaderboard screenshot of bad@0.5 for non-occluded pixels. Our model ranks first.



Coverage: dense | Set: Test | Metric: bad 1.0 [%] | Mask: non-occluded | Download this table as CSV

Method	Info	all	lakes. 1l	lakes. 1s	sand box 1l	sand box 1s	stora. room 1l	stora. room 1s	stora. room 2l	stora. room 2s	stora. room 2 1l	stora. room 2 1s	stora. room 2 2l	stora. room 2 2s	stora. room 3l	stora. room 3s	tunnel 1l	tunnel 1s	tunne 2
NBS		**0.14** 1	0.22 87	1.22 70	0.03 44	0.08 107	**0.08** 1	0.02 94	**0.36** 1	0.13 2	**0.02** 1	0.03 28	0.10 116	**0.03** 1	0.46 34	0.09 9	**0.00** 1	**0.00** 1	**0.0**
mm2		0.18 2	0.17 29	0.61 3	0.01 3	0.09 129	0.40 16	**0.00** 1	0.95 33	0.70 34	0.07 13	0.03 28	0.06 30	0.08 45	0.31 4	0.13 36	**0.00** 1	0.01 267	**0.0**
BLMT-Stereo		0.18 2	0.21 77	0.86 20	0.07 138	0.02 7	0.09 2	0.01 34	0.47 7	0.79 82	0.04 7	0.43 167	**0.02** 1	0.04 2	**0.28** 1	0.25 128	0.01 384	0.01 267	**0.0**
VeloStereo		0.20 4	0.19 50	**0.59** 1	0.01 3	0.08 107	1.14 72	**0.00** 1	0.60 13	0.39 17	0.10 19	0.04 46	0.07 55	0.11 88	0.37 16	0.18 90	**0.00** 1	0.01 267	**0.0**
GeoAtt-L (ZS)		0.21 5	0.32 210	1.02 32	0.09 175	0.12 181	0.29 9	0.05 152	0.58 11	0.19 5	0.15 50	0.19 130	0.07 55	0.05 10	0.82 125	0.13 36	0.01 384	0.01 267	**0.0**
CR-Stereo		0.22 6	0.20 64	1.06 36	0.02 20	0.02 7	0.71 29	0.02 94	0.38 3	0.70 34	0.12 26	**0.02** 1	0.07 55	0.05 10	0.95 143	0.11 21	**0.00** 1	0.01 267	**0.0**
DF_test		0.22 6	0.34 246	1.19 60	0.07 138	0.22 299	0.17 7	**0.00** 1	0.79 26	0.33 11	0.03 4	**0.02** 1	0.16 157	0.23 163	0.68 90	0.17 76	**0.00** 1	0.01 267	**0.0**
S2M2_XL		0.22 6	0.33 230	1.17 57	0.07 138	0.11 160	0.11 3	**0.00** 1	0.60 13	0.25 9	0.12 26	0.73 199	0.07 55	0.17 139	0.39 21	0.20 100	**0.00** 1	0.01 267	**0.0**
Junhong Min, Youngpil Jeon, Jimin Kim, Minyong Choi: S^2M^2 : Scalable Stereo Matching Model for Reliable Depth Estimation. ICCV 2025																			
LACA_dic		0.23 9	0.20 64	1.14 53	0.01 3	0.03 20	1.18 90	0.01 34	**0.36** 1	0.17 3	0.16 63	**0.02** 1	0.14 147	0.05 10	0.98 149	0.12 24	**0.00** 1	0.01 267	**0.0**

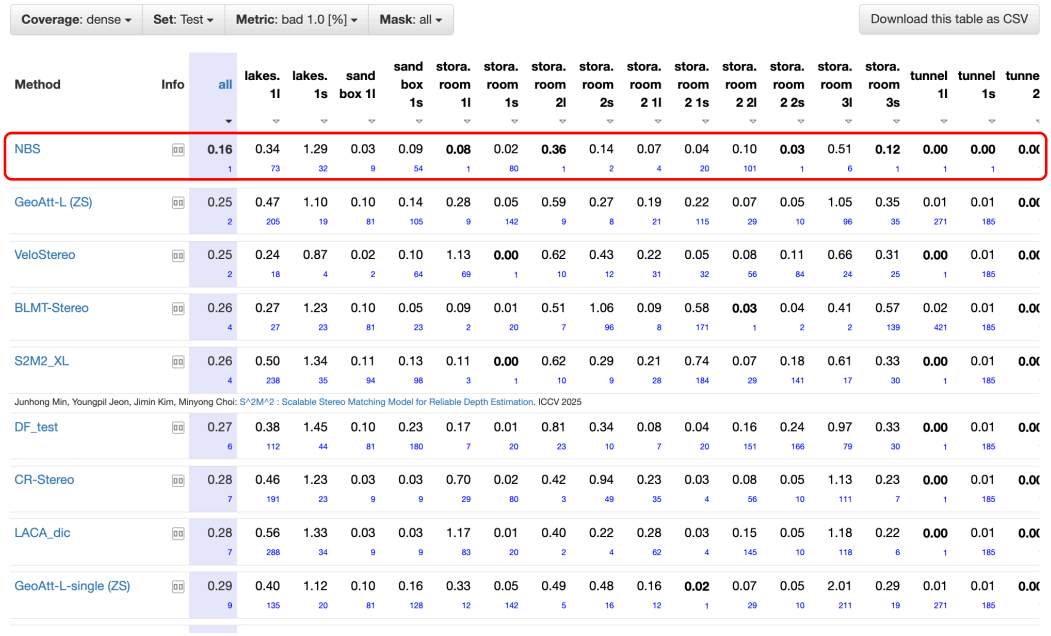
Figure 11: ETH3D leaderboard screenshot of bad@1 for non-occluded pixels. Our model ranks first.



Coverage: dense | Set: Test | Metric: bad 0.5 [%] | Mask: all | Download this table as CSV

Method	Info	all	lakes. 1l	lakes. 1s	sand box 1l	sand box 1s	stora. room 1l	stora. room 1s	stora. room 2l	stora. room 2s	stora. room 2 1l	stora. room 2 1s	stora. room 2 2l	stora. room 2 2s	stora. room 3l	stora. room 3s	tunnel 1l	tunnel 1s	tunne 2
NBS		**0.71** 1	0.80 34	3.16 17	2.82 90	0.33 23	**0.23** 1	0.05 34	**0.81** 1	**0.45** 1	**0.35** 1	1.17 63	0.28 12	**0.09** 1	3.13 27	**0.33** 1	0.02 51	0.02 60	**0.0**
GeoAtt-L (ZS)		0.75 2	0.74 14	**2.02** 1	**0.82** 1	0.37 40	0.36 2	0.10 57	1.18 3	1.10 2	1.46 56	0.86 51	**0.12** 1	0.10 2	4.87 103	0.89 43	0.04 120	0.01 21	**0.0**
GeoAtt-L-single (ZS)		0.93 3	0.78 26	2.34 4	1.97 9	0.45 52	0.52 5	0.08 42	1.05 2	1.16 4	1.89 107	0.34 20	0.44 40	0.26 10	6.31 152	0.94 54	0.02 51	0.01 21	**0.0**
WAFT-Stereo (ZS)		0.97 4	0.88 50	2.73 7	1.23 2	0.34 30	1.42 24	0.13 69	3.29 20	1.62 17	1.75 93	1.00 54	0.16 2	0.20 8	3.81 57	0.76 27	0.01 10	0.01 21	0.0 30
S2M2_XL		1.03 5	1.01 88	3.78 66	2.42 18	0.31 17	0.43 4	**0.01** 1	1.74 7	1.12 3	1.28 36	2.82 127	1.72 132	1.19 107	**1.48** 1	1.21 91	0.03 96	0.01 21	0.0 30
Junhong Min, Youngpil Jeon, Jimin Kim, Minyong Choi: S^2M^2 : Scalable Stereo Matching Model for Reliable Depth Estimation. ICCV 2025																			
Cross2Single-S		1.04 6	1.18 147	2.91 11	2.50 37	0.79 128	0.68 8	0.08 42	1.48 5	1.89 25	1.60 79	1.61 84	0.42 37	0.55 46	3.79 52	0.79 30	0.09 289	0.12 239	**0.0**
VeloStereo		1.14 7	0.82 40	3.56 42	2.42 18	0.39 43	1.50 26	0.04 28	2.29 12	1.94 30	0.98 16	0.62 35	2.03 157	0.97 88	3.96 60	1.13 81	0.02 51	0.03 81	**0.0**
DepthFocus		1.14 7	0.79 28	3.57 47	2.42 18	0.48 68	1.25 18	0.85 179	2.10 8	1.50 8	0.83 8	1.80 91	1.73 133	0.90 79	3.73 49	0.63 12	0.03 96	0.01 21	**0.0**
BLMT-Stereo		1.15 9	0.61 2	3.56 42	3.13 124	0.42 49	0.56 7	0.27 105	2.13 9	2.64 65	1.13 21	2.75 122	1.17 96	0.15 4	2.25 2	1.81 173	0.16 368	0.01 21	**0.0**

Figure 12: ETH3D leaderboard screenshot of bad@0.5 for all pixels. Our model ranks first.



Coverage: dense | Set: Test | Metric: bad 1.0 [%] | Mask: all | Download this table as CSV

Method	Info	all	lakes. 1l	lakes. 1s	sand box 1l	sand box 1s	stora. room 1l	stora. room 1s	stora. room 2l	stora. room 2s	stora. room 2 1l	stora. room 2 1s	stora. room 2 2l	stora. room 2 2s	stora. room 3l	stora. room 3s	tunnel 1l	tunnel 1s	tunne 2
NBS		**0.16** 1	0.34 73	1.29 32	0.03 9	0.09 54	**0.08** 1	0.02 80	**0.36** 1	0.14 2	0.07 4	0.04 20	0.10 101	**0.03** 1	0.51 6	**0.12** 1	**0.00** 1	**0.00** 1	**0.0**
GeoAtt-L (ZS)		0.25 2	0.47 205	1.10 19	0.10 81	0.14 105	0.28 9	0.05 142	0.59 9	0.27 8	0.19 21	0.22 115	0.07 29	0.05 10	1.05 96	0.35 35	0.01 271	0.01 185	**0.0**
VeloStereo		0.25 2	0.24 18	0.87 4	0.02 2	0.10 64	1.13 69	**0.00** 1	0.62 10	0.43 12	0.22 31	0.05 32	0.08 56	0.11 84	0.66 24	0.31 25	**0.00** 1	0.01 185	**0.0**
BLMT-Stereo		0.26 4	0.27 27	1.23 23	0.10 81	0.05 23	0.09 2	0.01 20	0.51 7	1.06 96	0.09 8	0.58 171	**0.03** 1	0.04 2	0.41 2	0.57 139	0.02 421	0.01 185	**0.0**
S2M2_XL		0.26 4	0.50 238	1.34 35	0.11 94	0.13 98	0.11 3	**0.00** 1	0.62 10	0.29 9	0.21 28	0.74 184	0.07 29	0.18 141	0.61 17	0.33 30	**0.00** 1	0.01 185	**0.0**
Junhong Min, Youngpil Jeon, Jimin Kim, Minyong Choi: S^2M^2 : Scalable Stereo Matching Model for Reliable Depth Estimation. ICCV 2025																			
DF_test		0.27 6	0.38 112	1.45 44	0.10 81	0.23 180	0.17 7	0.01 20	0.81 23	0.34 10	0.08 7	0.04 20	0.16 151	0.24 166	0.97 79	0.33 30	**0.00** 1	0.01 185	**0.0**
CR-Stereo		0.28 7	0.46 191	1.23 23	0.03 9	0.03 9	0.70 29	0.02 80	0.42 3	0.94 49	0.23 35	0.03 4	0.08 56	0.05 10	1.13 111	0.23 7	**0.00** 1	0.01 185	**0.0**
LACA_dic		0.28 7	0.56 288	1.33 34	0.03 9	0.03 9	1.17 83	0.01 20	0.40 2	0.22 4	0.28 62	0.03 4	0.15 145	0.05 10	1.18 118	0.22 6	**0.00** 1	0.01 185	**0.0**
GeoAtt-L-single (ZS)		0.29 9	0.40 135	1.12 20	0.10 81	0.16 128	0.33 12	0.05 142	0.49 5	0.48 16	0.16 12	**0.02** 1	0.07 29	0.05 10	2.01 211	0.29 19	0.01 271	0.01 185	**0.0**

Figure 13: ETH3D leaderboard screenshot of bad@1 for all pixels. Our model ranks first.