BENCHMARKING STEREO GEOMETRY ESTIMATION IN THE WILD

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the recent progress on stereo geometry estimation in the wild. Although recent stereo methods have achieved impressive benchmark results, the standard stereo benchmarks invest tremendous efforts into obtaining high-quality and perfectly calibrated stereo pairs, which is difficult to obtain in practice from in-the-wild settings. To address this in-the-wild evaluation gap, we introduce StereoBench, a benchmark for stereo, monocular, and multi-view geometry estimation methods comprising 26 method variants and 10 datasets within a unified depth-based evaluation protocol. Our findings reveal that although stereo methods perform well on high-quality benchmarks and out-of-domain synthetic data, they perform quite poorly on real-world data: even monocular methods with access to strictly less information often do better. We find that classic approaches for online stereo rectification cannot address this gap: instead, a far more effective strategy is to repurpose feed-forward multi-view geometry networks (such as VGGT) for calibration-robust stereo prediction, significantly outperforming dedicated stereo networks in the real world. We hope that our benchmark reveals crucial insights on robust stereo depth estimation in the wild that generalizes outside the domain of high-quality synthetic and benchmark inputs.

1 INTRODUCTION

Robust sensing is a crucial component required to scale robots to truly widespread deployments. A classic approach to depth sensing is stereopsis, motivated largely by stereo vision for human perception. Indeed, stereo may be the most benchmarked of all vision tasks, driven by iconic efforts such as Middlebury (Scharstein & Szeliski, 2002), KITTI (Geiger et al., 2012; Menze & Geiger, 2015), and ETH3D (Schops et al., 2017), all of which are now close to saturation. However, reported stereo performance on benchmarks does not seem to apply to the real world. Indeed, most autonomous vehicle efforts have disregarded stereo sensing after initial explorations (Chang et al., 2019). Why? We argue that classic benchmarks invested tremendous efforts into acquiring high-quality and perfectly calibrated stereo pairs, but these are difficult to acquire in practice. This begs the question of how one *should* benchmark stereo?

One reason for the disconnect is that stereo is often evaluated as a pixel-level disparity estimation task. While this elegantly *factors out* calibration parameters, such as extrinsics (the camera baseline) and intrinsics (the focal length), this also makes such evaluations unaware of sensitivities to calibration errors. Furthermore, it is not clear how to even define ground-truth disparity for input stereo pairs that are not perfectly rectified. Consider a stereo pair with imperfect undistortion or rectification, perhaps obtained from the Internet (Jin et al., 2025) or a real-world stereo setup. Should ground-truth disparity follow the ground-truth geometry, or pixel correspondences across scanlines in both images? Due to the overwhelming popularity of high-quality stereo benchmarks such as Middlebury and KITTI, we argue that these questions have been under-explored.

In this work, we propose to evaluate stereo geometry estimation on diverse synthetic and real-world datasets. We develop a unified evaluation protocol that measures errors in *scene depth rather than pixel disparity*. Though a straightforward benchmark design choice, this enables for the first time (to our knowledge) a fair comparison between stereo, monocular, and multi-view geometry estimation methods. We run our benchmark on 18 methods and 10 synthetic+real datasets spanning multiple diverse domains. One of our surprising conclusions is that monocular methods can out-

perform stereo on real data, even though they have access to strictly less information (one image versus two). One explanation is that calibration-free monocular methods implicitly predict camera intrinsics, while stereo methods explicitly factor out such parameters, which can hurt performance for real-world deployments with unreliable calibration parameters. Interestingly, classic approaches for online stereo rectification cannot address this gap. Instead, we find a far more effective strategy is to repurpose feedforward multi-view geometry networks (such as VGGT (Wang et al., 2025a) or Pi3 (Wang et al., 2025e)) for calibration-robust stereo prediction, which dramatically outperforms dedicated stereo engines in the real-world. Overall, our findings suggest the importance of developing robust in-the-wild stereo geometry estimators that generalize well outside the domain of perfectly-calibrated synthetic and benchmark inputs.

2 Related Work

2.1 Deep Stereo Matching

Modern advances in deep stereo matching are largely driven by three complementary factors: deep neural architectures, foundation model priors, and large-scale synthetic training data. Deep neural architectures, such as cost volumes (Yang et al., 2019; Xu & Zhang, 2020; Shen et al., 2021; 2022) and iterative refinement (Teed & Deng, 2020; Lipson et al., 2021; Li et al., 2022) have significantly enhanced the accuracy and generalization of stereo matching. Cost volume methods construct cost volumes from per-image features and use a 3D CNN for volume filtering, although the cost volume can be expensive. Iterative refinement methods perform recurrent disparity updates, avoiding expensive cost volumes but lacking global reasoning. Recent works therefore combine the strengths of both paradigms (Xu et al., 2023a). Another line of work introduces vision transformers to stereo matching, replacing domain-specific cost volumes with attention (Li et al., 2021; Weinzaepfel et al., 2023). Finally, with the rapid innovations in foundation models (Oquab et al., 2023) and monocular depth estimation (Yang et al., 2024b; Wang et al., 2025c; Piccinelli et al., 2025) several concurrent works introduce monocular priors to stereo matching (Bartolomei et al., 2025; Wen et al., 2025; Cheng et al., 2025) to provide informative geometry features as well as an initial disparity guess in ambiguous regions.

2.2 Stereo Datasets and Benchmarks

Training data and benchmarking is essential for deep learning models. Due to the perfect and dense ground-truth and the ease of data scaling, the dominant approach is to train deep stereo networks on synthetic data, such as TartanAir (Wang et al., 2020), Dynamic Replica (Karaev et al., 2023), and FSD (Wen et al., 2025). For evaluation, the standard benchmarks such as KITTI (Geiger et al., 2012; Menze & Geiger, 2015), Middlebury (Scharstein et al., 2014), and ETH3D (Schops et al., 2017) provide a small number of extremely high-quality real-world stereo pairs depicting typical indoor, outdoor, and driving scenarios with perfect calibration. (Guo et al., 2023) perform a comprehensive stereo matching benchmarking using the standard evaluation metrics. Recently, many larger real-world stereo datasets have recently been released, such as BotanicGarden (Liu et al., 2024), MS2 (Shin et al., 2023), SYNS (Adams et al., 2016), UASOL (Bauer et al., 2019), and VB-Rome (Brizi et al., 2024). These datasets typically capture highly accurate ground-truth via LiDAR or a laser scanner, making them suitable for benchmarking.

2.3 Monocular and Video Depth Estimation

Monocular depth estimation is a long-standing and fundamental task in computer vision. Many methods (Bhat et al., 2021; 2023; Eigen et al., 2014; Piccinelli et al., 2024; Hu et al., 2024; Yin et al., 2023) estimate depth with metric scale, supervised using data from RGB-D cameras, LiDAR sensors, and calibrated stereo cameras. Other methods predict depth up to a shift and scale (Li & Snavely, 2018; Ranftl et al., 2020; Yang et al., 2024a;b), allowing them to leverage a much wider variety of data, such as reconstructions from structure-from-motion algorithms (Li & Snavely, 2018). MiDaS (Ranftl et al., 2020), in particular, was the first work to popularize mixing large datasets from various domains for training. Recently, Marigold (Ke et al., 2024) shows the importance of using high-quality synthetic data for fine-grained predictions. DepthAnything (Yang et al., 2024a;b) fine-tunes foundational priors from DINOv2 while scaling up both real and synthetic data. MoGe

Table 1: **Datasets used in StereoBench.** StereoBench contains ˜20k frames taken from hundreds of real and synthetic videos, all with high-quality ground-truth. We report both the total each of images in the dataset, and the number of sampled frames for StereoBench evaluation. (1): Synthetic data. (2): Real data.

	Dataset	Total # Images	Sampled # Sequences	Sampled # Images	Avg Sampled Video Length	Domain
(1)	Eden	46455	369	4793	13	Nature Scenes
	IRS	26599	8	2533	345	Indoor Homes
	MidAir	26369	20	2700	138	Landscapes
	TartanAirV2	13961	8	1686	211	Indoor+Outdoor
	UnrealStereo4k	8200	9	820	91	Urban
(2)	BotanicGarden	1397	32	1397	44	Nature Scenes
	KITTI	1605	45	1605	36	Driving
	MS2	1631	20	1631	82	Driving (Night+Rain)
	SYNS	1244	74	1244	17	Outdoor+Nature
	VB-Rome	1181	8	1181	148	Urban

(Wang et al., 2025c) achieves high-quality geometry prediction via careful data curation, optimal alignment, and loss design. Simultaneously, several works explore video depth estimation, either by scaling up DINOv2-based monocular architectures (Chen et al., 2025), fine-tuning video diffusion backbones (Zhu et al., 2025), or combining video diffusion with monocular priors (Xu et al., 2025).

2.4 Multi-View Geometry Estimation

Rather than designing specific methods for distinct reconstruction tasks, such as monocular and stereo depth estimation, recent works have shown great promise in solving all reconstruction tasks jointly with a single feed-forward architecture. Enabled by recent advances in deep learning, recent methods like DUSt3R (Wang et al., 2024b), VGGSfM (Wang et al., 2024a), and VGGT (Wang et al., 2025a) can accept multiple unposed input images, performing feed-forward reconstruction by scaling up transformers on large datasets. In particular, DUSt3R (Wang et al., 2024b) and its metric-scale followup MASt3R (Leroy et al., 2024) take two unposed images as input and predict intrinsics, extrinsics, and scene geometry as a coupled pointmap representation. Recent works extend this paradigm with a memory mechanism (Wang & Agapito, 2025; Wang et al., 2025b; Cabon et al., 2025). Building on top of VGGT (Wang et al., 2025a), which modifies the DUSt3R architecture to support multi-view input and output, Pi3 (Wang et al., 2025e) removes the need for a fixed reference view, and MapAnything (Keetha et al., 2025) allows conditioning on optional input modalities.

3 StereoBench

Geometry estimation methods have historically fallen into different categories, such as monocular, stereo, multi-view, and video depth estimation, each with their own standard baselines, datasets, and evaluation metrics. As a result, previous literature lacks a comprehensive and fair comparison between stereo depth estimation and other categories of feed-forward geometry prediction, particularly on messy real-world data. To address this, we introduce a detailed benchmark called StereoBench: given an input stereo RGB video from one of 10 diverse datasets, we predict the per-frame depth map of the video up to scale. We describe the datasets, method, and evaluation metrics below.

3.1 Datasets

As shown in Fig. 1, StereoBench includes 10 diverse real and synthetic datasets spanning diverse domains, each with high-quality ground-truth. For real datasets, we include BotanicGarden (Liu et al., 2024), KITTI 2012 (Geiger et al., 2012), MS2 (Shin et al., 2023), SYNS (Adams et al., 2016), and VB-Rome (Brizi et al., 2024). Collectively, these span diverse environments such as natural foliage, autonomous driving, driving in unique weather conditions, as well as both urban

Figure 1: **Sample images from datasets used in StereoBench.** StereoBench contains diverse scenery representative of common stereo use cases, including nature scenes, driving, urban areas, landscapes, and indoor homes.

and rural scenes across multiple countries. To ensure accurate ground-truth, all real-world datasets are captured with stereo RGB alongside high-quality LiDAR or laser scanners (e.g. Ouster OS0-128, Velodyne HDL-64E, Velodyne VLP16). For synthetic datasets, we include Eden (Le et al., 2021), IRS (Wang et al., 2021), MidAir (Fonder & Droogenbroeck, 2019), TartanAirV2 (Wang et al., 2020), and UnrealStereo4k (Tosi et al., 2021). These datasets span indoor homes, gardens, natural landscapes, and synthetic structures such as supermarkets, factories, and caves. Due to the large size of synthetic datasets, we randomly select about 10% of the videos as the test split, with several thousand images per dataset (see Tab. 1 for more details). Compared to standard benchmarks such as Middlebury (Scharstein et al., 2014) and ETH3D (Schops et al., 2017), our evaluation is much larger and contains more diverse real and synthetic environments.

3.2 Methods

We evaluate 18 methods across three categories of feed-forward geometry prediction methods: monocular depth, stereo depth, and video or multi-view.

1. We include six state-of-the-art monocular depth models spanning major design choices: Depth Anything V2 (Yang et al., 2024b), Depth Pro (Bochkovskii et al., 2025), Lotus (He et al., 2025), Metric3Dv2 (Hu et al., 2024), MoGeV2 (Wang et al., 2025d), and UniDepthV2 (Piccinelli et al., 2025). These include both diffusion and DINOv2-based models, affine-invariant and metric-scale predictions, as well as models that focus mainly on sharp geometry prediction vs. accurate relative or metric-scale depth.
2. We include six SOTA stereo depth models: CREStereo (Li et al., 2022), FoundationStereo (Wen et al., 2025), DUSt3R (Wang et al., 2024b), MASt3R (Leroy et al., 2024), UFM (Zhang et al., 2025), and UniMatch (Xu et al., 2023b). Of these, the former two are designed for pure stereo depth estimation given perfect stereo pairs. DUSt3R and MASt3R are designed for unconstrained stereo depth estimation given unposed images. The latter two are designed for optical flow: we use the horizontal flow for stereo evaluation.
3. Finally, we include six multi-view geometry models: VGGT (Wang et al., 2025a), Pi3 (Wang et al., 2025e), MapAnything (Keetha et al., 2025), Metric Video Depth Anything (Chen et al., 2025), Aether (Zhu et al., 2025), and GeometryCrafter (Xu et al., 2025). Of these, the former three perform unconstrained multi-view geometry estimation given unposed images, while the latter three perform video depth estimation.

3.3 Evaluation Metrics

Monocular, stereo, video and multi-view geometry estimation methods typically use different evaluation metrics, making it hard to compare results across domains. Monocular methods typically report relative depth error (AbsRel), RMSE, and inlier rates (typically at a 25% threshold), often after performing a scale or scale+shift alignment between predictions and ground-truth. Stereo methods

typically report disparity metrics, such as endpoint error and inlier rates at various pixel or relative thresholds. Multi-view methods often report chamfer distance metrics (Wang et al., 2025a).

To accommodate all three method categories, our benchmark will report standard depth metrics for all three method categories following (Eigen et al., 2014), including AbsRel error ($\frac{1}{|T|}\sum_{y\in|T|}\frac{|y-y*|}{y*}$) and inlier rate at a 1.25 threshold ($\%$ of y_i s.t. $\max(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}) < 1.25$). We consider two main tasks: monocular depth estimation and video depth estimation. For both tasks, we perform scale alignment between the predictions and ground-truth, following the optimal alignment procedure in (Wang et al., 2025c). All metrics are computed over videos, then averaged across the videos in a dataset. To assess the quality of metric scale prediction, we also report the average Scale Ratio produced by alignment of predictions to ground-truth, in a similar manner as AbsRel: $\max(s, \frac{1}{s})$ where s minimizes $\frac{1}{|T|}\sum_{y\in T}\frac{|sy-y*|}{y*}$. Note that the minimum value of Scale Ratio is 1, if no scale alignment is needed. Finally, we report the average rank of each method according to the δ_1 metric. Our code and results will be publicly released.

To obtain additional insights on how the same model performs on different tasks, we also evaluate multi-view and video methods on monocular and stereo depth estimation where appropriate. For multi-view models such as VGGT, we run N separate monocular or stereo inferences on individual frames, rather than passing the entire N-frame video at once. For video models, we simply replicate individual frames to generate a static video.

4 Experiments

We conduct empirical experiments on StereoBench to assess the quality of stereo, monocular, and multi-view geometry prediction. Our experiments aim to address the following research questions:

Q1: How do stereo depth methods and other method categories perform when given real vs. synthetic data, and when computing per-frame depth vs. video depth?

Q2: How do geometric foundation models perform on tasks they were not designed for, compared to domain-specific models? For example, can VGGT and Pi3 effectively serve as monocular and stereo depth estimators, even though they were not specifically trained for this task?

Q3: How does stereo rectification impact in-the-wild performance of stereo depth estimation?

4.1 Relative Single-Frame Geometry Estimation

Task Formulation

We first aim to assess the relative single-frame geometry estimation capabilities for various methods and datasets. Given a single stereo pair, we input each left frame separately to monocular methods, input both frames to stereo methods, and provide the entire left video as input to multi-view and video methods (using a sliding window approach for Aether (Zhu et al., 2025) and GeometryCrafter (Xu et al., 2025)). We follow (Wang et al., 2025c) (Sec. A.2) and find an optimal per-frame scale factor between depth predictions and ground-truth: this scale factor is also reported. However, for methods which require affine-invariant disparity alignment (such as DAv2 (Yang et al., 2024b) and Lotus (He et al., 2025)), we use the least-squares solver for disparity scale and shift following VGGSfM (Wang et al., 2024a).

Results

Tab. 2 shows our results on synthetic data, and Tab. 3 shows our results on real data. We share several key takeaways below:

(1) Stereo depth estimators perform well on synthetic data, but very poorly on real data. In Tab. 2 and 3, stereo methods (FoundationStereo and CREStereo) rank on top for almost every synthetic dataset, but are ranked on the bottom for every real dataset (except KITTI, which is a standard stereo benchmark known for its high quality). There is a significant synthetic-to-real gap in stereo geometry estimation (over 4x larger error for FoundationStereo: 6.56% $\rightarrow$ 28.76% AbsRel and 95.08% $\rightarrow$ 60.25% δ_1 error). Although other method categories also suffer from a synthetic-to-

Table 2: Relative single-frame geometry estimation on synthetic data. We report Rel^p (↓), δ_1^p (↑), and Scale Ratio (↓), where Rel^p and δ_1^p are percentages.

Method	Rank↓	Average			Eden			IRS			MidAir			TartanAirV2			UnrealStereo4k		
		Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓
DepthAnythingV2	11.40	13.28	88.27	N/A	18.91	82.29	N/A	10.65	92.67	N/A	10.37	90.83	N/A	17.80	82.80	N/A	8.65	92.78	N/A
Depth Pro	18.00	21.33	73.78	5.33	13.84	80.15	1.86	11.15	89.95	1.56	12.02	83.71	3.25	20.02	74.81	2.06	49.63	40.27	17.93
Lotus	19.40	17.16	82.04	N/A	20.57	79.69	N/A	12.60	88.29	N/A	14.38	83.08	N/A	27.15	70.27	N/A	11.11	88.88	N/A
Metric3Dv2	19.00	13.25	82.36	1.72	16.19	76.11	1.97	9.98	89.69	1.74	13.89	80.54	2.17	16.68	76.63	1.33	9.51	88.83	1.39
MoGeV2	6.20	7.67	92.49	1.48	9.92	90.65	2.83	6.37	94.41	1.12	4.92	96.96	1.06	13.23	83.31	1.34	3.90	97.13	1.05
UniDepthV2	14.40	13.63	86.34	2.04	8.70	92.57	4.10	9.92	90.84	1.34	20.13	79.89	1.88	22.57	74.06	1.75	6.82	94.35	1.13
DUST3R-Mono	24.20	19.61	70.36	13.07	23.04	63.56	3.47	15.05	78.32	4.25	20.99	66.12	37.55	25.58	62.21	10.24	13.41	81.58	9.83
MASt3R-Mono	21.20	15.42	78.36	1.79	19.59	70.27	1.23	13.21	83.03	1.13	16.93	74.55	4.21	21.19	69.12	1.35	6.21	94.82	1.05
MapAnything-Mono	16.00	11.50	85.47	1.64	15.26	78.18	3.18	11.04	87.20	1.38	11.53	85.00	1.44	13.81	82.13	1.13	5.86	94.83	1.09
Pi3-Mono	14.00	10.09	87.89	8.69	13.54	81.28	2.05	9.37	88.98	2.01	6.52	94.98	29.49	14.31	80.77	5.20	6.73	93.43	4.71
VGGT-Mono	9.60	8.27	90.90	9.65	8.53	92.03	2.15	9.72	88.39	2.42	4.27	97.32	30.10	15.08	79.58	7.29	3.73	97.19	6.28
CREStereo	5.60	6.74	93.59	**1.03**	9.06	93.92	1.02	8.11	91.29	**1.05**	4.03	96.36	**1.00**	9.27	89.76	1.04	3.25	96.64	1.02
FoundationStereo	**2.00**	**6.56**	**95.08**	1.04	12.56	**95.52**	**1.01**	8.18	92.00	1.14	**2.96**	**97.51**	1.01	**6.71**	**92.86**	**1.03**	**2.39**	**97.50**	**1.02**
DUSt3R	20.60	13.90	81.52	15.38	16.27	76.84	4.02	11.09	87.57	4.49	12.30	83.66	46.56	20.37	70.74	11.51	9.50	88.81	10.30
MASt3R	9.80	9.37	90.52	1.92	9.82	90.89	1.19	9.95	89.35	1.13	8.75	91.89	4.78	13.89	83.91	1.45	4.47	96.55	1.04
UFM	16.00	14.68	86.33	1.06	13.06	89.67	1.02	15.57	84.54	1.17	14.07	87.17	1.01	19.31	81.14	1.05	11.40	89.15	1.03
UniMatch	10.20	11.25	88.47	1.35	8.41	94.09	1.01	26.88	71.14	2.64	6.29	93.44	1.01	10.58	87.86	1.06	4.07	95.80	1.04
MapAnything-Stereo	13.40	10.81	86.72	1.65	14.39	79.59	3.21	10.47	88.38	1.39	10.45	87.01	1.44	13.19	83.43	1.13	5.53	95.19	1.10
Pi3-Stereo	5.60	6.94	93.32	8.76	7.63	93.26	2.11	7.34	92.56	2.00	4.73	96.64	29.71	10.16	88.32	5.33	4.82	95.85	4.64
VGGT-Stereo	4.00	6.77	93.39	9.71	**7.05**	94.42	2.17	8.02	91.65	2.45	3.86	97.48	30.14	11.46	86.04	7.52	3.45	97.35	6.25
Aether	25.60	46.44	40.30	61.17	77.67	14.42	192.73	26.66	54.89	18.14	33.44	49.80	2.50	43.05	36.16	13.80	51.39	46.21	78.66
GeometryCrafter	16.60	21.29	76.90	4.45	15.50	78.65	2.45	8.86	91.99	nan	60.72	39.31	nan	14.10	81.24	5.54	7.27	93.33	5.34
MapAnything-Video	13.40	11.04	86.49	1.64	14.70	79.48	3.18	9.65	90.38	1.31	11.32	85.47	1.48	13.65	82.28	1.14	5.86	94.83	1.09
Pi3-Video	8.60	8.24	91.41	26.91	10.22	88.42	4.46	6.86	93.17	3.74	6.05	95.61	101.92	11.33	86.42	19.72	6.73	93.43	4.71
VGGT-Video	5.20	7.07	93.00	49.38	7.15	93.83	4.53	**6.11**	**94.76**	3.70	4.62	97.09	206.09	13.75	82.12	26.29	3.73	97.19	6.28
VideoDA-Metric	20.40	22.09	73.66	2.33	49.46	45.20	5.52	10.39	90.38	1.43	17.30	75.70	1.49	20.93	71.21	1.58	12.40	85.84	1.63

Table 3: Relative single-frame geometry estimation on real data. We report Rel^p (↓), δ_1^p (↑), and Scale Ratio (↓), where Rel^p and δ_1^p are percentages.

Method	Rank↓	Average			BotanicGarden			KITTI			MS2			SYNS			VB-Rome		
		Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓
DepthAnythingV2	12.20	18.13	81.05	N/A	29.17	70.84	N/A	10.71	89.73	N/A	20.33	74.52	N/A	14.94	84.71	N/A	15.52	85.45	N/A
Depth Pro	16.40	15.93	79.22	1.45	22.50	67.65	1.81	8.45	91.40	1.10	18.74	74.23	1.20	14.67	80.73	1.88	15.32	82.11	1.26
Lotus	17.80	17.57	78.76	N/A	23.78	67.93	N/A	11.87	86.42	N/A	20.73	73.18	N/A	15.42	82.98	N/A	16.07	83.31	N/A
Metric3Dv2	4.80	13.19	84.49	1.33	18.64	75.94	1.58	6.81	93.71	1.09	**14.93**	**82.91**	1.11	13.09	82.63	1.68	12.49	87.24	1.19
MoGeV2	4.80	13.46	83.84	1.29	18.83	73.61	**1.13**	7.14	93.60	1.29	16.65	79.73	1.14	12.64	85.09	1.67	12.05	87.18	1.20
UniDepthV2	3.20	14.33	84.58	**1.23**	22.28	73.96	1.59	7.33	94.00	1.05	15.75	82.49	1.11	12.67	**85.86**	**1.20**	13.64	86.60	1.23
DUST3R-Mono	22.40	18.89	73.40	26.01	26.24	61.28	12.85	11.96	83.54	27.04	20.86	70.13	29.24	17.87	74.62	33.77	17.52	77.43	27.16
MASt3R-Mono	16.60	16.28	78.79	2.89	22.25	69.54	1.88	10.42	87.56	2.41	17.96	75.02	2.29	15.86	78.91	5.16	14.93	82.90	2.71
MapAnything-Mono	11.00	16.26	81.40	1.32	24.59	69.83	1.49	9.05	91.11	1.08	19.45	77.41	1.08	13.31	83.65	1.57	14.90	85.01	1.37
Pi3-Mono	10.00	14.66	82.05	12.10	19.25	73.62	5.12	8.89	90.90	14.11	19.45	76.30	15.42	12.92	83.48	13.03	12.80	85.96	12.82
VGGT-Mono	18.60	15.38	77.69	14.51	19.88	68.22	6.44	9.17	88.97	18.89	17.93	73.56	16.97	14.93	77.54	16.46	14.99	80.15	13.77
CREStereo	16.40	26.60	62.95	1.39	32.52	57.86	1.32	6.58	94.20	1.02	19.82	76.90	**1.06**	28.64	53.21	1.83	45.45	32.58	1.74
FoundationStereo	17.00	28.76	60.25	1.58	35.11	55.23	1.59	5.92	**94.93**	1.01	21.03	75.77	1.10	38.38	39.42	2.65	43.36	35.90	1.55
DUSt3R	19.20	17.12	77.42	27.01	22.83	68.98	13.52	10.19	87.23	27.71	19.59	73.13	30.23	16.41	77.03	35.06	16.56	80.72	28.52
MASt3R	5.80	13.71	84.63	2.91	18.75	78.92	1.96	7.39	92.45	2.61	15.38	81.33	2.33	13.15	85.01	5.27	13.86	85.44	2.38
UFM	24.20	38.91	52.86	16.42	47.97	49.16	1.17	14.44	86.71	1.01	24.37	71.97	1.15	38.59	38.19	2.63	69.20	18.29	76.15
UniMatch	14.80	29.90	60.14	16.79	21.74	69.67	1.16	**5.48**	94.55	**1.01**	19.12	77.55	1.08	37.63	38.95	2.56	65.52	19.97	78.12
MapAnything-Stereo	8.20	15.64	82.52	1.30	23.13	72.77	1.48	8.54	91.88	1.07	19.00	78.15	1.07	13.18	83.86	1.55	14.34	85.93	1.34
Pi3-Stereo	**3.00**	**12.73**	**85.81**	12.12	**15.96**	**80.09**	5.07	7.26	93.37	14.01	17.36	80.41	15.52	**11.80**	85.71	13.09	**11.29**	**89.46**	12.91
VGGT-Stereo	13.60	14.30	80.26	14.66	17.78	72.88	6.57	8.36	90.55	18.92	16.89	76.03	17.05	14.38	78.56	16.68	14.10	83.28	14.06
Aether	25.20	42.24	41.95	70.10	53.72	27.48	3.24	21.02	63.23	2.97	36.84	40.83	1.87	69.76	25.73	340.48	29.89	52.45	1.93
GeometryCrafter	8.40	14.98	82.36	16.44	22.40	70.99	6.46	7.66	93.12	19.73	17.82	78.20	18.74	14.11	83.20	20.36	12.91	86.30	16.89
MapAnything-Video	12.40	16.30	81.07	1.34	24.35	69.54	1.41	8.67	91.65	1.11	20.22	75.65	1.15	13.31	83.65	1.57	14.98	84.84	1.46
Pi3-Video	6.80	13.86	83.26	41.27	18.51	75.02	25.23	7.07	93.27	40.05	17.94	78.14	70.79	12.92	83.48	13.03	12.86	86.40	57.25
VGGT-Video	15.20	14.36	79.00	86.37	18.91	70.22	37.85	8.44	90.48	63.31	16.47	76.60	201.27	14.93	77.54	16.46	13.05	80.15	112.94
VideoDA-Metric	22.40	23.02	69.64	1.54	38.76	45.74	2.25	12.32	86.81	1.19	20.86	71.43	1.40	24.81	66.73	1.69	18.36	77.50	**1.17**

real drop in performance, it is much less severe (e.g. for MoGeV2: 7.67% → 13.46% AbsRel and 92.49% → 83.84% δ_1 error).

Table 4: Metric-scale video geometry estimation on synthetic data. Rel^p and δ_1^p are percentages.

Method	Average				Eden			IRS			MidAir			TartanAirV2			UnrealStereo4k		
	Rank↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓
DepthAnythingV2	19.20	54.41	48.32	N/A	79.41	39.15	N/A	72.92	36.24	N/A	41.44	38.82	N/A	69.57	34.67	N/A	8.70	92.71	N/A
Depth Pro	19.40	47.57	44.88	6.25	79.12	31.05	1.77	30.08	72.29	5.91	45.50	27.76	2.10	33.18	53.31	3.46	49.98	40.00	18.00
Lotus	19.80	55.10	47.33	N/A	71.74	42.23	N/A	76.69	33.82	N/A	38.94	40.38	N/A	77.06	31.36	N/A	11.09	88.85	N/A
Metric3Dv2	14.60	23.53	62.47	1.70	35.24	44.90	2.03	14.45	83.16	1.72	31.30	39.73	1.96	27.16	55.74	1.42	9.51	88.84	1.39
MoGeV2	5.80	12.87	82.83	1.48	21.72	63.09	2.85	9.88	92.21	1.10	6.43	96.68	1.06	22.41	65.05	1.32	3.90	97.12	1.05
UniDepthV2	9.60	18.99	79.93	2.15	11.00	89.98	4.10	17.57	80.18	1.29	22.48	80.81	2.19	37.01	54.36	2.06	6.90	94.32	1.13
DUST3R-Mono	24.20	42.69	38.43	11.26	66.19	29.50	3.43	41.63	27.98	3.28	40.79	30.46	32.57	51.46	22.66	7.16	13.41	81.58	9.83
MASt3R-Mono	17.80	31.33	52.20	1.71	46.52	39.41	1.40	29.47	51.63	1.21	37.67	33.23	3.54	36.77	41.89	1.34	6.21	94.82	1.05
MapAnything-Mono	12.60	22.65	64.27	1.67	36.58	42.34	3.37	22.88	65.16	1.36	27.34	46.94	1.36	20.61	72.06	1.15	5.86	94.83	1.09
Pi3-Mono	17.00	30.09	50.65	7.51	39.19	45.21	2.05	34.98	36.32	1.68	29.97	44.11	25.15	39.60	34.19	3.95	6.73	93.42	4.71
VGGT-Mono	14.80	30.40	50.95	8.36	38.10	46.72	2.09	39.54	31.21	1.92	28.16	47.61	26.12	42.46	32.00	5.41	3.73	97.19	6.28
CREStereo	4.00	7.32	93.77	1.02	10.24	93.32	1.03	9.25	91.85	**1.02**	4.03	96.55	**1.00**	9.82	90.52	1.00	3.25	96.64	1.02
FoundationStereo	**1.00**	**7.07**	**95.54**	**1.01**	12.65	**95.81**	**1.00**	10.79	**93.06**	1.03	**2.83**	**97.79**	1.00	**6.69**	**93.55**	**1.00**	**2.39**	**97.50**	**1.02**
DUSt3R	22.40	37.89	43.48	13.35	57.78	34.67	3.80	39.26	30.72	3.60	35.53	37.17	40.75	47.38	26.03	8.30	9.50	88.81	10.30
MASt3R	13.80	27.09	57.70	1.74	37.09	47.87	1.34	27.62	57.97	1.18	34.40	38.34	3.80	31.89	47.75	1.32	4.47	96.55	1.04
UFM	9.20	16.39	85.49	1.08	13.86	89.92	1.01	22.26	79.53	1.37	14.09	87.19	1.01	20.35	81.65	1.01	11.40	89.15	1.03
UniMatch	7.00	13.72	86.51	1.61	9.52	93.43	1.05	37.02	60.64	3.97	6.47	93.60	1.00	11.55	89.07	1.00	4.07	95.81	1.04
MapAnything-Stereo	11.20	22.00	65.36	1.67	35.55	43.43	3.40	22.37	66.60	1.37	26.46	48.49	1.36	20.08	73.10	1.15	5.53	95.19	1.10
Pi3-Stereo	13.40	28.22	53.15	7.55	34.92	50.13	2.04	34.50	37.99	1.70	29.40	45.26	25.23	37.45	36.53	4.12	4.82	95.85	4.64
VGGT-Stereo	13.60	29.49	51.99	8.40	37.09	47.01	2.09	38.58	33.29	2.00	27.91	47.63	26.06	40.42	34.67	5.60	3.45	97.34	6.25
Aether	25.80	71.89	17.62	904.83	100.59	2.78	4202.80	61.56	18.51	89.10	62.83	14.99	5.98	81.84	6.74	126.91	52.62	45.08	99.34
GeometryCrafter	16.20	33.58	55.92	4.31	33.14	48.79	2.42	27.30	61.40	nan	71.61	25.05	nan	28.60	51.04	5.15	7.27	93.32	5.34
MapAnything-Video	8.60	18.74	72.24	1.65	32.02	49.53	3.32	14.34	84.31	1.27	22.60	57.79	1.43	18.89	74.74	1.14	5.86	94.83	1.09
Pi3-Video	6.80	10.83	87.59	26.87	15.30	77.73	4.42	**8.09**	92.54	3.74	7.98	94.68	101.54	16.06	79.58	19.96	6.73	93.42	4.71
VGGT-Video	4.40	9.58	89.20	49.28	**8.62**	92.52	4.55	8.70	91.96	3.75	5.64	97.07	205.58	21.23	67.25	26.24	3.73	97.19	6.28
VideoDA-Metric	18.20	32.55	56.44	2.36	69.37	29.19	5.72	13.64	86.32	1.42	35.04	33.38	1.37	32.33	47.48	1.67	12.40	85.82	1.63

(2) Geometric foundation models do not suffer from a large synthetic-to-real quality gap. Indeed, Pi3-Stereo is the best-performing model on average in Tab. 3, outperforming even dedicated approaches such as MASt3R. Pi3-Stereo improves from among the top 3 methods to the best-performing method on average after switching from synthetic to real data (Rank 5.40 → Rank 2.80). Among other geometric foundation models, although VGGT-Stereo has a significant synthetic-to-real domain gap (going from Rank 4.00 with synthetic data to Rank 13.60 with real data), the gap is still smaller than stereo methods. MapAnything-Stereo actually improves in the rankings from synthetic to real (Rank 13.40 to Rank 8.20).

(3) Domain-specific models outperform multi-view models repurposed for specific tasks. With the exception of multi-view models repurposed for stereo, where stereo depth estimators suffer from a substantial synthetic-to-real gap, the monocular versions of multi-view models such as Pi3 and VGGT perform slightly worse than domain-specific depth models (in Tab. 3: Rank 10.00 for Pi3-Mono compared to Rank 3.20 for UniDepthV2). Similarly, for synthetic data in Tab. 2, Pi3Stereo (Rakn 5.60) and VGGTStereo (Rank 4.00) are on par with CREStereo (Rank 5.60), but have not yet reached the quality of FoundationStereo (Rank 2.00). However, the gap seems to be closing as models improve.

4.2 Metric-Scale Video Geometry Estimation

Task Formulation

Beyond per-image relative depth estimation, we also aim to evaluate metric-scale video geometry estimation, which is more often the desired task in reconstruction pipelines. Given a stereo video, we input each left frame separately to monocular methods, input both frames to stereo methods, and provide the entire left video as input to multi-view and video methods. As before, we follow (Wang et al., 2025c) to find an optimal per-video scale factor between depth predictions and ground-truth, and use alternative affine disparity alignments for DAv2 and Lotus.

Results

Tab. 4 shows our results on synthetic data, and Tab. 5 shows our results on real data. We share several key takeaways below:

Table 5: Metric-scale video geometry estimation on real data. Rel^p and δ_1^p are percentages.

Method	Rank↓	Average			BotanicGarden			KITTI			MS2			SYNS			VB-Rome		
		Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓
DepthAnythingV2	15.40	29.70	61.85	N/A	51.85	40.71	N/A	17.84	75.79	N/A	28.51	55.05	N/A	15.42	84.53	N/A	34.88	53.19	N/A
Depth Pro	16.00	22.92	64.98	1.42	38.43	37.66	1.74	12.41	85.78	1.11	21.11	68.67	1.18	14.67	80.80	1.90	27.99	52.01	1.18
Lotus	14.20	25.95	65.78	N/A	30.53	57.48	N/A	19.07	71.94	N/A	31.42	58.69	N/A	15.51	82.96	N/A	33.24	57.85	N/A
Metric3Dv2	**4.00**	**14.19**	**82.98**	1.31	**19.19**	**75.35**	1.50	7.34	93.50	1.09	**15.11**	**82.79**	1.11	13.08	82.62	1.68	16.24	**80.65**	**1.16**
MoGeV2	4.20	16.02	80.10	1.29	22.62	66.12	**1.14**	8.65	92.99	1.29	18.48	76.92	1.15	12.64	85.09	1.66	17.68	79.35	1.21
UniDepthV2	4.20	18.04	77.59	**1.25**	27.74	60.15	1.63	8.48	93.42	1.05	16.10	82.12	1.10	12.68	**85.79**	**1.20**	25.18	66.49	1.28
DUST3R-Mono	19.40	26.30	58.43	24.91	38.11	39.15	12.74	15.69	77.12	26.82	26.91	53.67	28.02	17.87	74.61	33.77	32.93	47.62	23.18
MASt3R-Mono	20.80	31.32	48.44	2.63	33.99	46.06	1.82	30.34	47.51	2.15	37.98	34.08	1.88	15.86	78.92	5.16	38.45	35.64	2.16
MapAnything-Mono	10.00	21.05	72.55	1.33	30.76	57.38	1.49	11.76	87.15	1.09	21.32	72.78	1.07	13.31	83.65	1.57	28.11	61.78	1.42
Pi3-Mono	17.00	24.75	60.58	11.13	32.35	43.36	4.75	14.18	81.03	13.88	27.92	51.76	14.34	12.92	83.47	13.03	36.38	43.26	9.65
VGGT-Mono	20.60	26.04	56.90	13.85	38.07	35.26	6.21	15.02	77.87	18.87	26.96	52.22	16.55	14.93	77.54	16.46	35.20	41.60	11.15
CREStereo	12.40	28.45	61.12	1.40	30.34	57.06	1.19	6.42	94.38	1.01	19.32	78.39	**1.04**	28.62	53.22	1.83	57.54	22.53	1.94
FoundationStereo	13.80	35.20	54.33	1.64	38.35	49.49	1.73	5.87	**95.00**	1.01	20.14	77.17	1.04	38.37	39.40	2.65	73.29	10.58	1.79
DUSt3R	17.20	24.73	61.48	25.88	35.13	43.95	13.24	14.17	80.94	27.46	25.64	55.96	28.82	16.41	77.02	35.05	32.30	49.52	24.85
MASt3R	17.20	29.32	53.96	2.69	30.91	53.93	1.91	28.59	52.31	2.36	37.88	37.60	1.97	13.15	85.03	5.27	36.08	40.95	1.95
UFM	18.40	40.16	51.17	22.60	48.13	48.78	1.18	14.47	86.68	1.01	26.65	69.06	1.21	38.60	38.22	2.63	72.97	13.11	106.97
UniMatch	11.20	30.82	59.23	23.16	22.07	69.62	1.19	**5.50**	94.54	**1.01**	18.52	78.76	1.04	37.59	39.17	2.55	70.44	14.05	110.02
MapAnything-Stereo	8.60	20.48	73.88	1.32	29.52	59.78	1.49	11.29	88.37	1.08	20.66	74.35	1.06	13.18	83.85	1.55	27.77	63.03	1.41
Pi3-Stereo	14.00	23.52	62.80	11.20	30.30	46.99	4.70	13.44	82.59	13.81	26.72	53.19	14.44	**11.79**	85.70	13.09	35.32	45.52	9.97
VGGT-Stereo	19.60	25.61	57.82	13.92	37.53	36.10	6.21	14.66	78.74	18.89	26.53	52.98	16.57	14.38	78.57	16.68	34.98	42.73	11.24
Aether	25.00	59.29	25.23	483.01	91.47	2.87	1974.70	23.33	57.86	3.13	55.79	21.48	10.19	71.98	23.94	421.87	53.89	20.02	5.14
GeometryCrafter	9.20	19.52	73.80	16.22	29.89	55.02	6.44	9.38	91.27	19.73	19.75	75.24	18.82	14.11	83.18	20.36	24.48	64.28	15.75
MapAnything-Video	8.00	20.01	74.48	1.34	28.80	61.26	1.40	10.45	89.63	1.10	21.92	71.12	1.13	13.31	83.65	1.57	25.58	66.75	1.48
Pi3-Video	5.80	15.95	79.49	41.28	21.63	69.68	25.39	7.99	92.85	40.02	18.83	76.58	71.11	12.92	83.47	13.03	18.40	74.86	56.83
VGGT-Video	8.80	16.08	76.86	87.20	22.11	65.25	38.94	9.86	89.87	63.83	18.00	73.85	201.67	14.93	77.54	16.46	**15.51**	77.80	115.09
VideoDA-Metric	15.40	25.15	65.22	1.51	41.13	39.43	2.05	12.93	85.94	1.19	21.27	70.20	1.36	24.82	66.69	1.69	25.59	63.83	1.26

(1) On real data, per-frame metric monocular depth outperforms other method categories. In Tab. 5, monocular methods such as Metric3Dv2, MoGeV2, and UniDepthV2 surprisingly outperform methods specifically designed for multi-view or video depth, even when the monocular methods are inferred separately per-frame. Metric3Dv2 and MoGeV2 achieve Rank 4.00 and 4.20, while Pi3-Video, VGGT-Video, and VideoDA-Metric achieve Rank 5.80, 8.80, and 15.40 respectively. This suggests that even when the goal is video depth estimation, monocular geometry estimation is a strong baseline. High-quality metric-scale video depth estimation remains a hard task, as even temporally-inconsistent per-frame monocular depth can outperform existing arts.

(2) A key advantage of stereo is perfect metric scale estimation given accurate calibration. Assuming high-quality stereo pairs with accurate intrinsics and stereo baseline, no other methods can come close to stereo at metric scale estimation: this can be seen for synthetic data in Tab. 4 and for high-quality real benchmarks such as KITTI in Tab. 5. The best data-driven method for monocular scale estimation on KITTI is UniDepthV2, and it still has a 5% scale error, compared to a 1% scale error for all stereo methods. However, the poor performance of stereo on many in-the-wild datasets suggests that this finding may not be useful in practice.

(3) Metric scale errors dominate over relative depth errors when evaluating video depth. For every method category (except stereo depth estimation on synthetic or high-quality benchmark data), the magnitude of metric scale errors is larger than the magnitude of relative errors, even when scale alignment is performed over the entire video. For example, in Tab. 5, UniDepthV2 has a 25% scale error on average, but a 18% relative depth error on average. This suggests that data-driven models lack an extremely precise understanding of metric scale.

4.3 Impact of Online Stereo Rectification

As stereo depth estimation requires rig-rectified and undistorted stereo pairs, there are an abundance of classical methods for offline stereo calibration (Loop & Zhang, 1999) as well as online stereo rectification given correspondences (Hansen et al., 2012; Hartley, 1999). To validate whether errors in stereo calibration (e.g. rig rectification and undistortion) are responsible for poor stereo geometry estimation in the real world, we preprocess all real-world stereo datasets with a rectifying homography algorithm to obtain rectified stereo pairs with horizontal scanlines. We then evaluate all stereo methods on these online-rectified stereo pairs.

Table 6: Metric-scale video geometry estimation on real and synthetic data for stereo methods, when online stereo rectification is applied. Rel^p and δ_1^p are percentages. Although online rectification can improve the performance of stereo algorithms such as CREStereo and FoundationStereo, the improved versions still lag behind per-frame monocular depth results in Tab. 5.

Method	Rank↓	Average			BotanicGarden			KITTI			MS2			SYNS			VB-Rome		
		Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓	Rel^p ↓	δ_1^p ↑	Scale↓
CREStereo	5.40	19.50	75.46	1.15	28.01	61.55	1.18	5.81	94.98	1.03	19.39	78.43	1.05	18.92	75.22	1.31	25.38	67.12	1.15
FoundationStereo	6.20	21.07	73.55	1.25	33.74	54.70	1.55	5.37	**95.45**	1.03	20.60	76.59	1.05	21.50	71.06	1.39	24.12	69.97	1.23
DUSt3R	16.20	26.27	57.65	23.74	35.75	41.68	12.56	14.90	78.77	28.36	25.67	55.67	27.95	21.16	67.73	26.10	33.88	44.41	23.71
MASt3R	13.20	28.53	54.41	2.32	30.09	54.64	1.94	33.21	44.03	2.06	35.31	41.04	1.85	13.06	83.15	3.75	31.00	49.19	1.99
UFM	10.60	28.89	68.35	1.73	45.87	50.58	1.18	13.07	88.00	1.03	25.67	71.41	1.24	21.28	70.00	1.35	38.56	61.78	3.86
UniMatch	5.60	19.95	75.94	1.70	22.19	69.73	1.18	5.54	94.45	1.04	18.61	78.63	1.06	19.43	72.79	1.34	34.00	64.10	3.86
MapAnything-Stereo	9.20	21.54	71.04	1.29	31.64	53.13	1.60	11.53	88.74	1.09	21.21	74.13	1.04	15.52	78.73	1.42	27.79	60.46	1.30
Pi3-Stereo	12.20	23.58	61.91	10.38	29.83	47.29	4.62	13.33	82.37	13.34	26.72	53.30	13.42	12.88	83.42	10.93	35.14	43.18	9.56
VGGT-Stereo	15.60	25.84	56.80	13.65	35.25	38.78	6.26	14.14	80.78	19.97	28.66	48.86	16.33	18.99	70.41	13.52	32.15	45.17	12.15

Implementation Details

We follow the approach in (Hansen et al., 2012) (Eq. 1-8), which computes an online rectifying homography $\mathbf{R}_L, \mathbf{R}_R$ given pairs of image correspondences. Here, $\mathbf{R}_L$ and $\mathbf{R}_R$ are rotation matrices which separately rotate the left and right camera coordinate systems, to ensure correct stereo rig calibration (i.e. zero relative rotation and only horizontal translation). Specifically, for each input video, we compute image correspondences using a state-of-the-art flow model (Zhang et al., 2025). As we assume that the stereo rig rectification is fixed across an input video, we concatenate optical flow correspondences across the entire video, and estimate a single $\mathbf{R}_L$ and $\mathbf{R}_R$ per video via non-linear least squares minimization of epipolar errors.

Results

Our results are available in Tab. 6. Compared to Tab. 5, performing performing online stereo rectification significantly increases the performance of CREStereo (average δ_1 from $61.12 \rightarrow 75.46$) and FoundationStereo (average δ_1 from $54.33 \rightarrow 73.55$). However, alternative methods such as Pi3-Video (average $\delta_1 = 79.49$) and Metric3Dv2 (average $\delta_1 = 82.98$) remain more competitive.

5 Conclusion

We introduce StereoBench, a benchmark for stereo, monocular, and multi-view geometry estimation comparing 18 methods across 10 diverse datasets within a unified evaluation protocol. We discover several key findings: (1) Stereo methods perform surprisingly poorly on real-world datasets, despite ranking above all other models on both high-quality benchmarks and out-of-domain synthetic data. (2) In contrast, data-driven methods such as monocular depth and multi-view geometric foundation models do not suffer from large synthetic-to-real gaps. On real data, per-frame monocular depth outperforms most other model categories at video depth prediction. (3) Although online stereo rectification can partially improve stereo performance on real-world data, it also worsens performance on well-calibrated datasets, and cannot fully account for stereo depth's synthetic-to-real gap. We will continue updating our benchmark to support all datasets, methods, and evaluation metrics that are of interest to the community. We hope that our benchmark inspires future research in robust real-world stereo depth estimation on practical in-the-wild data.

Reproducibility statement. All data, evaluation code, and benchmarks will be publicly released.

References

Wendy Adams, James Elder, Erich Graf, Julian Leyland, Arthur Lugtigheid, and Alexander Muryy. The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. *Scientific Reports*, 2016. URL https://syns.soton.ac.uk/.

Luca Bartolomei, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Stereo Anywhere: Robust Zero-Shot Deep Stereo Matching Even Where Either Stereo or Mono Fail. *CVPR*, 2025. URL https://arxiv.org/abs/2412.04472.

Zuria Bauer, Francisco Gomez-Donoso, Edmanuel Cruz, Sergio Orts-Escolano, and Miguel Cazorla. UASOL and a large-scale high-resolution outdoor stereo dataset. *Scientific Data*, 2019. URL https://www.rovit.ua.es/dataset/uasol.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. AdaBins: Depth Estimation using Adaptive Bins. *CVPR*, 2021. URL https://arxiv.org/abs/2011.14141.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. *arXiv*, 2023. URL https://arxiv.org/abs/2302.12288.

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. *ICLR*, 2025. URL https://arxiv.org/abs/2410.02073.

Leonardo Brizi, Emanuele Giacomini, Luca Di Giammarino, Simone Ferrari, Omar Salem, Lorenzo De Rebotti, and Giorgio Grisetti. VBR: A Vision Benchmark in Rome. *ICRA*, 2024. URL https://arxiv.org/abs/2404.11322.

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. MUSt3R: Multi-view Network for Stereo 3D Reconstruction. *CVPR*, 2025. URL https://arxiv.org/abs/2503.01661.

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3D Tracking and Forecasting with Rich Maps. *CVPR*, 2019. URL https://arxiv.org/abs/1911.02620.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video Depth Anything: Consistent Depth Estimation for Super-Long Videos. *CVPR*, 2025. URL https://arxiv.org/abs/2501.12375.

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. *CVPR*, 2025. URL https://arxiv.org/abs/2501.08643.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. *NeurIPS*, 2014. URL https://arxiv.org/abs/1406.2283.

Michael Fonder and Marc Van Droogenbroeck. Mid-Air: A multi-modal dataset for extremely low altitude drone flights. *CVPR-W*, 2019.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. *CVPR*, 2012.

Xianda Guo, Chenming Zhang, Juntao Lu, Yiqun Duan, Yiqi Wang, Tian Yang, Zheng Zhu, and Long Chen. OpenStereo: A Comprehensive Benchmark for Stereo Matching and Strong Baseline. *arXiv*, 2023. URL https://arxiv.org/abs/2312.00343.

Peter Hansen, Hatem Alismail, Peter Rander, and Brett Browning. Online continuous stereo extrinsic parameter estimation. *CVPR*, 2012.

Richard I. Hartley. Theory and Practice of Projective Rectification. *IJCV*, 1999.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based Visual Foundation Model for High-quality Dense Prediction. *ICLR*, 2025. URL https://arxiv.org/abs/2409.18124.

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. *TPAMI*, 2024. URL https://arxiv.org/abs/2404.15506.

Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski. Stereo4D: Learning How Things Move in 3D from Internet Stereo Videos. *CVPR*, 2025. URL https://arxiv.org/abs/2412.09621.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent Dynamic Depth from Stereo Videos. *CVPR*, 2023. URL https://arxiv.org/abs/2305.02296.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. *CVPR*, 2024. URL https://arxiv.org/abs/2312.02145.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal Feed-Forward Metric 3D Reconstruction. *arXiv*, 2025. URL https://www.arxiv.org/abs/2509.13414.

Hoang-An Le, Thomas Mensink, Partha Das, Sezer Karaoglu, and Theo Gevers. EDEN: Multimodal Synthetic Dataset of Enclosed GarDEN Scenes. *WACV*, 2021. URL https://arxiv.org/abs/2011.04389.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding Image Matching in 3D with MASt3R. *arXiv*, 2024. URL https://arxiv.org/abs/2406.09756.

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. *CVPR*, 2022. URL https://arxiv.org/abs/2203.11483.

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. *ICCV*, 2021. URL https://arxiv.org/abs/2011.02910.

Zhengqi Li and Noah Snavely. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. *CVPR*, 2018. URL https://arxiv.org/abs/1804.00607.

Lahav Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. *3DV*, 2021. URL https://arxiv.org/abs/2109.07547.

Yuanzhi Liu, Yujia Fu, Minghui Qin, Yufeng Xu, Baoxin Xu, Fengdong Chen, Bart Goossens, Poly Z.H. Sun, Hongwei Yu, Chun Liu, Long Chen, Wei Tao, and Hui Zhao. BotanicGarden: A High-Quality Dataset for Robot Navigation in Unstructured Natural Environments. *RA-L*, 2024. URL https://arxiv.org/abs/2306.14137.

Charles Loop and Zhengyou Zhang. Computing Rectifying Homographies for Stereo Vision. *CVPR*, 1999.

Moritz Menze and Andreas Geiger. Object Scene Flow for Autonomous Vehicles. *CVPR*, 2015.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *arXiv*, 2023. URL https://arxiv.org/abs/2304.07193.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal Monocular Metric Depth Estimation. *CVPR*, 2024. URL https://arxiv.org/abs/2403.18913.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal Monocular Metric Depth Estimation Made Simpler. *arXiv*, 2025. URL https://arxiv.org/abs/2502.20110.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *TPAMI*, 2020. URL https://arxiv.org/abs/1907.01341.

Daniel Scharstein and Richard Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *IJCV*, 47(1):7–42, 2002. URL https://vision.middlebury.edu/stereo/taxonomy-IJCV.pdf.

Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nesic, Xi Wang, and Porter Westling. High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. *CVPR*, 2014.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. *CVPR*, 2017.

Zhelun Shen, Yuchao Dai, and Zhibo Rao. CFNet: Cascade and Fused Cost Volume for Robust Stereo Matching. *CVPR*, 2021. URL https://arxiv.org/abs/2104.04314.

Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. PCW-Net: Pyramid Combination and Warping Cost Volume for Stereo Matching. *ECCV*, 2022. URL https://arxiv.org/abs/2006.12797.

Ukcheol Shin, Jinsun Park, and In So Kweon. Multi-Spectral Stereo (MS2) Outdoor Driving Dataset. *CVPR*, 2023.

Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. *ECCV*, 2020. URL https://arxiv.org/abs/2003.12039.

Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. SMD-Nets: Stereo Mixture Density Networks. *CVPR*, 2021. URL https://arxiv.org/abs/2104.03866.

Hengyi Wang and Lourdes Agapito. Spann3R: 3D Reconstruction with Spatial Memory. *3DV*, 2025. URL https://arxiv.org/abs/2408.16061.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGSfM: Visual Geometry Grounded Deep Structure From Motion. *CVPR*, 2024a. URL https://arxiv.org/abs/2312.04563.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. *CVPR*, 2025a. URL https://arxiv.org/abs/2503.11651.

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. *ICME*, 2021. URL https://arxiv.org/abs/1912.09678.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D Perception Model with Persistent State. *CVPR*, 2025b. URL https://arxiv.org/abs/2501.12387.

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. *CVPR*, 2025c. URL https://arxiv.org/abs/2410.19115.

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. *arXiv*, 2025d. URL https://arxiv.org/abs/2507.02546.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. *CVPR*, 2024b. URL https://arxiv.org/abs/2312.14132.

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. *IROS*, 2020. URL https://arxiv.org/abs/2003.14338.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. Pi3: Permutation-Equivariant Visual Geometry Learning. *arXiv*, 2025e. URL https://arxiv.org/abs/2507.13347.

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. *ICCV*, 2023. URL https://arxiv.org/abs/2211.10408.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-Shot Stereo Matching. *CVPR*, 2025. URL https://arxiv.org/abs/2501.09898.

Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative Geometry Encoding Volume for Stereo Matching. *CVPR*, 2023a. URL https://arxiv.org/abs/2303.06615.

Haofei Xu and Juyong Zhang. AANet: Adaptive Aggregation Network for Efficient Stereo Matching. *CVPR*, 2020. URL https://arxiv.org/abs/2004.09548.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying Flow and Stereo and Depth Estimation. *TPAMI*, 2023b. URL https://arxiv.org/abs/2211.05783.

Tian-Xing Xu, Xiangjun Gao, Wenbo Hu, Xiaoyu Li, Song-Hai Zhang, and Ying Shan. GeometryCrafter: Consistent Geometry Estimation for Open-world Videos with Diffusion Priors. *ICCV*, 2025. URL https://arxiv.org/abs/2504.01016.

Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical Deep Stereo Matching on High-Resolution Images. *CVPR*, 2019. URL https://arxiv.org/abs/1912.06704.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. *CVPR*, 2024a. URL https://arxiv.org/abs/2401.10891.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. *NeurIPS*, 2024b. URL https://arxiv.org/abs/2406.09414.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image. *ICCV*, 2023.

Yuchen Zhang, Nikhil Keetha, Chenwei Lyu, Bhuvan Jhamb, Yutian Chen, Yuheng Qiu, Jay Karhade, Shreyas Jha, Yaoyu Hu, Deva Ramanan, Sebastian Scherer, and Wenshan Wang. UFM: A Simple Path towards Unified Dense Correspondence with Flow. *arXiv*, 2025. URL https://arxiv.org/abs/2506.09278.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-Aware Unified World Modeling. *ICCV*, 2025. URL https://arxiv.org/abs/2503.18945.