

StereoDiff: Stereo-Diffusion Synergy for Video Depth Estimation

Supplementary Material

Region	AbsRel ↓	RMSE ↓	$\delta 1 \uparrow$	$\delta 2 \uparrow$
Dynamic	-0.0463	-0.5809	+0.0982	+0.0294
Overall	-0.0191	+0.2364	+0.0375	-0.0017
Static	-0.0171	+0.2968	+0.0326	-0.0042

(a) Performance improvement of StereoDiff over MonST3R.

Region	AbsRel ↓	RMSE ↓	$\delta 1 \uparrow$	$\delta 2 \uparrow$
Dynamic	+0.0110	-0.4692	-0.0344	-0.0131
Overall	-0.0147	-0.9638	+0.0067	+0.0128
Static	-0.0184	-1.0070	+0.0126	+0.0163

(b) Performance improvement of StereoDiff over DepthCrafter.

Table 1. **Quantitative comparisons on dynamic and static regions of KITTI**, following the settings in main paper’s Tab. 3.

Method	Bonn	KITTI	ScanNetV2	Sintel	Avg. Rank
DepthAnything V2	0.522	2.052	0.627	1.421	7.0
DepthAnything	0.510	1.899	0.613	1.463	6.5
DUST3R	0.546	2.273	0.491	2.838	7.8
MASt3R	0.532	2.126	0.536	2.537	7.8
MonST3R _{OPT}	—	1.766	—	2.241	5.0
MonST3R	<u>0.439</u>	1.823	0.507	2.342	4.5
ChronoDepth	0.507	1.894	0.583	1.579	6.0
DepthCrafter	0.489	1.780	0.552	1.139	3.8
DepthAnyVideo	0.474	<u>1.694</u>	0.531	<u>1.380</u>	<u>2.8</u>
StereoDiff (Ours)	0.387	1.595	0.470	1.389	1.5

Table 2. **Quantitative comparisons on temporal consistency**. StereoDiff delivers the *lowest* avg. rank, demonstrating its superior temporal consistency. Please see Sec. 3 for the specific process.

Method	DepthCrafter	MonST3R	StereoDiff (Ours)
Inf. Time (s)	1.1708	0.4100	0.4100+0.1569

Table 3. **Inference time per frame** tested on the first scene of Bonn dataset (“balloon”), using an NVIDIA A800 GPU. We set $n = 2$ for both MonST3R and StereoDiff.

1. Qualitative Comparisons

Qualitative comparisons on four *in-the-wild* (or zero-shot), dynamic, and read-world video depth benchmarks, among DepthCrafter [2], MonST3R [13], and StereoDiff are illustrated in Fig. 1, 2, and 3. In static regions, StereoDiff effectively utilizes stereo matching to deliver highly robust and stable video depth estimations. In dynamic regions, StereoDiff excels in maintaining smooth local consistency across consecutive frames, addressing challenges posed by both the object motion and camera movement.

Note that before visualization, *both predicted and GT depth maps are normalized by the maximum depth value of*

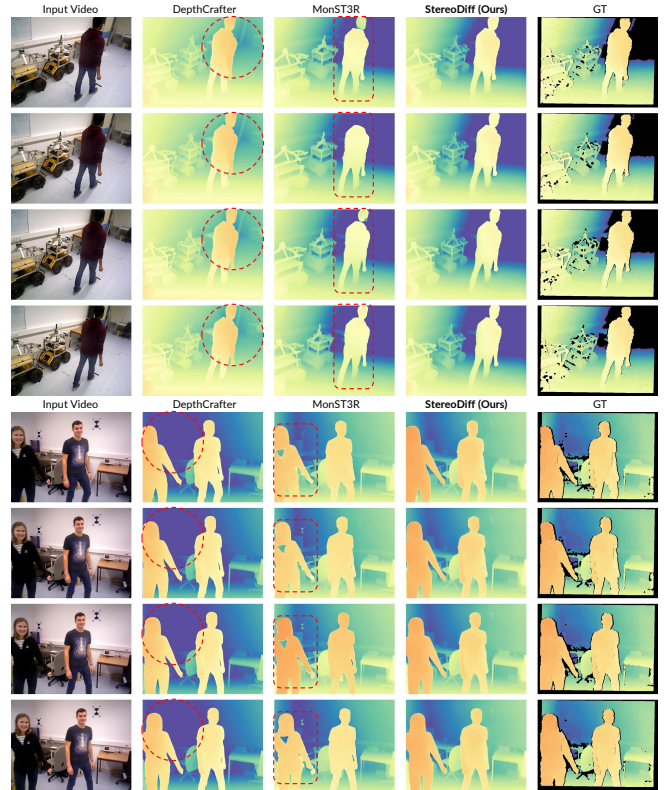


Figure 1. **Qualitative comparisons on Bonn dataset**, conducted among MonST3R, DepthCrafter, and StereoDiff. Four continuous frames are sampled from a video depth sequence to form a complete comparison set. Please visit the [project page](#) or the uploaded Supplementary Materials for video comparisons.

the evaluation dataset, which means that the visualization are plotted in *metric* scale rather than relative.

2. Frequency Analysis on 3D Trajectories

As illustrated in Fig. 4 and 5, compared with main paper’s Fig. 3, the error sequences on 3D trajectories even *more clearly* demonstrate StereoDiff’s effective synergy for the advantages of both stereo matching and video diffusion—StereoDiff delivers lower or comparable error in both low frequencies compared with MonST3R (especially in \bullet , \bullet), and also delivers lower or comparable error in high frequencies compared with DepthCrafter (especially in \bullet , \bullet).

3. Temporal Consistency

Following CVD [4], we report quantitative experiments on temporal consistency: 1) Use GT camera intrinsics to lift the

Metrics	Method	Low Freq.								High Freq.
		\mathcal{F}_0	\mathcal{F}_1	\mathcal{F}_2	\mathcal{F}_3	\mathcal{F}_4	\mathcal{F}_5	\mathcal{F}_6	\mathcal{F}_7	\mathcal{F}_8
AbsRel↓	DepthCrafter	0.1620	0.0306	0.0324	0.0363	0.0272	0.0169	0.0129	0.0103	0.0076
	MonST3R	0.1666	<u>0.0258</u>	<u>0.0221</u>	0.0277	0.0279	0.0208	0.0190	0.0135	0.0135
	StereoDiff (Ours)	0.1476	0.0209	0.0155	<u>0.0285</u>	0.0247	<u>0.0171</u>	<u>0.0136</u>	<u>0.0106</u>	<u>0.0078</u>
RMSE↓	DepthCrafter	5.4048	0.7941	0.8940	1.0056	0.8343	<u>0.4651</u>	0.3548	0.2641	0.1965
	MonST3R	4.1926	<u>0.4247</u>	<u>0.3956</u>	0.4656	<u>0.5366</u>	0.5599	0.5215	0.3529	0.2526
	StereoDiff (Ours)	<u>4.4291</u>	0.2985	0.3678	<u>0.5270</u>	0.5345	0.4628	<u>0.3496</u>	<u>0.2690</u>	<u>0.2293</u>
$(1 - \delta_1)\downarrow$	DepthCrafter	<u>0.2322</u>	0.0635	0.0671	0.0821	0.0674	0.0482	0.0352	0.0269	0.0204
	MonST3R	0.2647	0.0679	0.0506	0.0977	0.0853	<u>0.0605</u>	0.0555	0.0428	0.0427
	StereoDiff (Ours)	0.2304	<u>0.0777</u>	<u>0.0557</u>	<u>0.0930</u>	<u>0.0744</u>	0.0618	<u>0.0403</u>	<u>0.0325</u>	<u>0.0262</u>

Table 4. **Quantitative comparisons across different frequency domains on KITTI**, following the settings in main paper’s Tab. 2. The entire frequency range is grouped exponentially into 9 discrete bands, \mathcal{F}_0 to \mathcal{F}_8 , representing low to high frequencies.

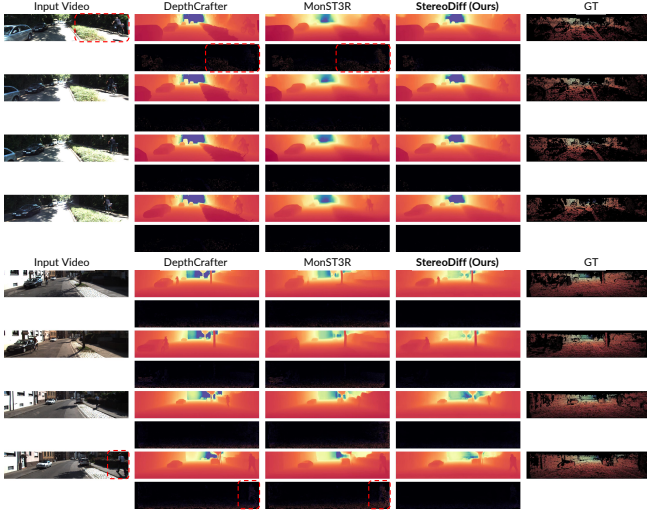


Figure 2. **Qualitative comparisons on KITTI dataset**. For better clarity, the corresponding error maps are provided below each estimated depth map. Please zoom in for detailed views.

predicted video depth maps \hat{D} into dynamic 3D points; 2) Use GT optical flows (only Sintel) or CoTracker3 for dense 2D flows prediction in *static* areas; 3) Project \hat{D}_i ’s 3D points to \hat{D}_j using GT camera poses (i, j : evenly, Δ -spaced frame-indexes, $i \neq j, \Delta=10$), and compute avg. Euclidean distance of point pairs. As shown Tab. 2, StereoDiff delivers the *lowest* avg. rank, showing its superior temporal consistency.

4. Inference Speed

The inference time comparison among MonST3R [13], DepthCrafter [2] and StereoDiff is reported in Tab. 3. Thanks to efficient stereo matching and MST alignment, especially the one-step denoising policy of the video depth diffusion model in the second stage, StereoDiff is ~ 2.1 times faster than DepthCrafter.

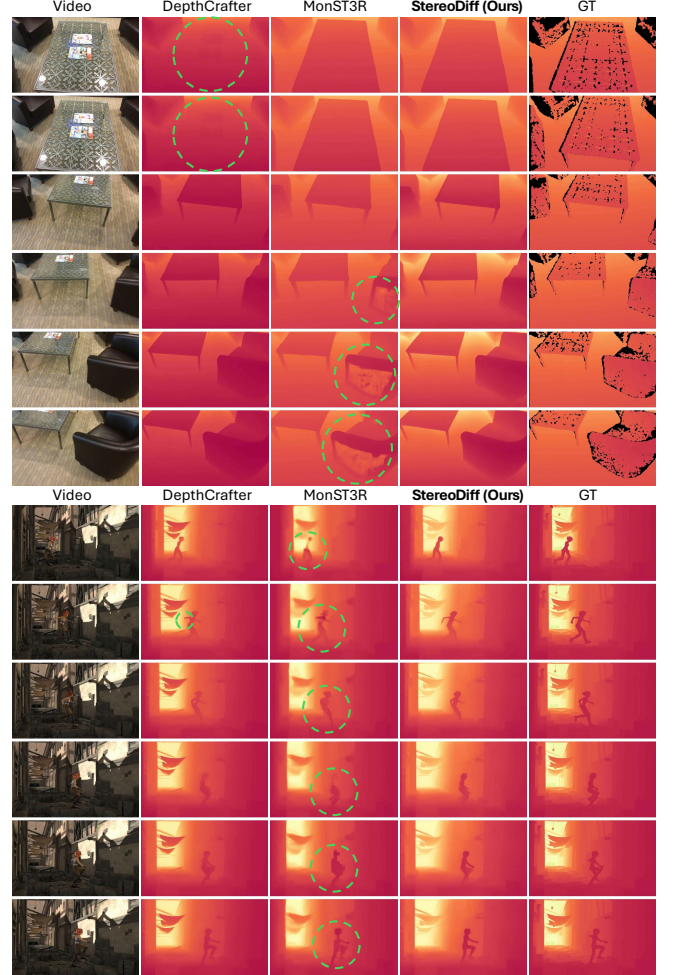


Figure 3. **Qualitative comparisons on ScanNetV2 and Sintel**. On ScanNetV2, StereoDiff shows clear superiority over DepthCrafter and MonST3R. On Sintel, StereoDiff is comparable with DepthCrafter and superior over MonST3R.



Figure 4. **3D trajectories on StereoDiff’s dynamic 3D points** using CoTracker3 [3]. 2 points are randomly sampled from static areas and 2 points on dynamic areas. Please zoom in for details.

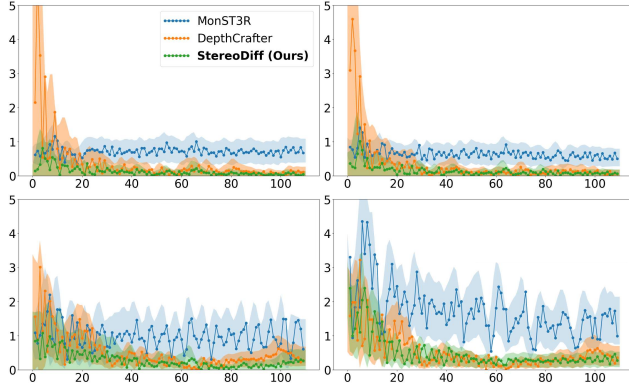


Figure 5. **Magnitude spectrum of the error sequence (Euclidean distance) on 3D trajectories** (X: Frequency (Hz); Y: Amplitude; From top left to bottom right: ●, ●, ●, ●). The settings are inherited from main paper’s Fig. 3, only the 3D trajectory-covered frames are utilized. Please zoom in for details.

5. Limitations

The limitation of StereoDiff mainly stems from its first stage, which is a stereo matching process designed to achieve robust and strong global consistency through global 3D constraints. SfM methods [1, 5–13] inevitably face failure cases due to various limitations. These include challenges with textureless or repetitive surfaces, constantly changing lighting conditions, and computational challenges in large-scale scenarios. While the second-stage of StereoDiff can significantly reduce deficiency, the various limitations cannot be entirely avoided.

References

- [1] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 1272–1279, 2013. 3
- [2] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 1, 2
- [3] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. 2024. 3
- [4] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (ToG)*, 39(4):71–1, 2020. 1
- [5] Kemal E Ozden, Konrad Schindler, and Luc Van Gool. Multi-body structure-from-motion in practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1134–1141, 2010. 3
- [6] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.
- [7] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2016.
- [8] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006.
- [9] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024.
- [10] 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*, 2024.
- [11] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022.
- [12] Changchang Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*, pages 127–134. IEEE, 2013.
- [13] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 1, 2, 3