Seurat: From Moving Points to Depth

Supplementary Material

Depth Estimator	Point Tracker	Aria [10]			DriveTrack [1]			PStudio [6]			**Average**		
		3D-AJ ↑	APD ↑	TC ↓	3D-AJ ↑	APD ↑	TC ↓	3D-AJ ↑	APD ↑	TC ↓	3D-AJ ↑	APD ↑	TC ↓
Oracle depth*	CoTracker [8]	55.9	70.3	-	53.2	71.7	-	46.9	65.0	-	52.0	69.0	-
ZoeDepth [2]	Oracle tracker*	11.8	18.6	0.04	8.6	13.9	1.22	14.2	22.2	0.05	11.5	18.2	0.44
	CoTracker [8]	9.8	15.8	0.06	7.2	**12.3**	1.26	10.2	17.9	0.05	9.1	15.3	0.46
	LocoTrack [4]	9.6	15.7	0.05	**7.5**	**12.3**	1.33	9.9	17.3	0.06	9.0	15.1	0.48
	SpatialTracker [13]	9.2	15.1	-	5.8	10.2	-	9.8	17.7	-	8.3	14.3	-
+ Seurat (Ours)	CoTracker	10.9	18.8	**0.01**	6.6	11.6	**0.27**	10.9	18.8	**0.01**	9.6	16.4	**0.10**
DepthPro [3]	Oracle tracker*	12.2	18.9	0.13	7.0	12.0	3.72	11.0	18.0	0.16	10.1	16.3	1.34
	CoTracker [8]	9.9	15.9	0.15	5.2	9.4	3.38	7.8	14.4	0.16	7.6	13.2	1.23
	LocoTrack [4]	9.2	15.6	0.13	5.3	9.3	3.60	7.8	14.2	0.17	7.4	13.0	1.30
+ Seurat (Ours)	CoTracker	**14.6**	**21.9**	**0.01**	6.9	11.8	**0.27**	**12.7**	**20.7**	**0.01**	**11.4**	**18.1**	**0.10**

Table A. **Quantitative results on TAPVid-3D [9] minival split with meidan scaling.** We combined the depth ratio from Seurat with the metric depth predictions from ZoeDepth [2] and DepthPro [3]. *Oracle tracker** rows use ground-truth 2D trajectories, while *Oracle depth** row uses ground-truth depth estimation to determine the upper bound.

The supplementary materials begin with an additional quantitative comparison in Sec. A. Next, we present an analysis of inference time in Sec. B. Sequentially, Sec. C provides additional implementation details. Finally, we discuss the limitations of our work in Sec. D.

A. More Results

Quantitative comparison with median scaling. In Table A, we compare our method with baselines that combine depth estimators [2, 3] and point trackers [4, 7]. Overall, our method outperforms the baselines, with particularly significant improvements over those using DepthPro. Additionally, our method demonstrates substantially better temporal coherency (TC), further highlighting its effectiveness in maintaining consistent depth predictions over time.

Quantitative comparison using depth metrics. In Table B and Table C, we compare our method with other baselines using metrics widely adopted in depth estimation literature [5, 11, 12, 14]. Specifically, we employ the absolute relative error (AbsRel) and δ_1. AbsRel is calculated as $|\hat{d} - d|/d$, while δ_1 is defined as the percentage of $\max\left(\hat{d}/d, d/\hat{d}\right) < 1.25$, where $\hat{d}$ denotes the predicted depth and d denotes the ground-truth depth. In Table B, since the compared depth estimators [5, 12] predict affine-invariant depth, we apply scale-and-shift optimization using least squares to align their predicted depth scales with the ground truth. For Seurat that uses the trajectory of CoTracker [8] as input consistently outperforms other baselines significantly, validating its effectiveness in depth accuracy.

B. More Analysis

Analysis on inference time. Table D presents an analysis of inference time with different numbers of query points. We separately measure the inference time required for point tracking and depth inference using our method. The results show that point tracking accounts for most of the computation time, while our model is relatively efficient. We believe that future advancements in point tracking efficiency will lead to more efficient inference for our overall pipeline.

C. More Implementation Details

During training, we sample $N_q = 256$ query trajectories per batch. While we utilize trajectories from off-the-shelf models [4, 8] during inference, we use ground-truth trajectory positions and occlusion information as input during training. Occluded positions in the input trajectories are masked by replacing their values with the last visible position. In addition to the depth prediction head at the end of our model, we include a head to predict the position of occluded points, using the same loss function as in [8]. We found this beneficial for the model to produce smooth depth estimates for occluded points. For iterative depth refinement, we consistently use 4 iterations for both training and inference.

D. Limitations and Discussion

We have shown that the temporal evolution of depth can be inferred from trajectories extracted by off-the-shelf point trackers. However, our model has a limited ability to infer spatial relative depth, relying instead on the monocular depth estimation model. End-to-end training of this combined pipeline could further synergize temporal and spa-

Depth Estimator	Point Tracker	Aria [10] AbsRel ↓	Aria [10] δ_1 ↑	DriveTrack [1] AbsRel ↓	DriveTrack [1] δ_1 ↑	PStudio [6] AbsRel ↓	PStudio [6] δ_1 ↑	**Average** AbsRel ↓	**Average** δ_1 ↑
DepthCrafter [5]	Oracle tracker*	0.344	0.564	0.141	0.811	0.053	0.981	0.179	0.785
	CoTracker [8]	0.390	0.542	0.165	0.787	0.057	0.974	0.204	0.768
	LocoTrack [4]	0.394	0.529	0.182	0.784	0.058	0.973	0.211	0.762
ChronoDepth [12]	Oracle tracker*	0.248	0.671	0.106	0.887	0.064	0.977	0.139	0.845
	CoTracker [8]	0.287	0.644	0.132	0.843	0.067	0.971	0.162	0.819
	LocoTrack [4]	0.290	0.648	0.166	0.811	0.068	0.971	0.175	0.810
Seurat (Ours)	CoTracker [8]	**0.198**	**0.754**	**0.127**	**0.868**	**0.052**	**0.984**	**0.126**	**0.869**
+ ZoeDepth [2]	LocoTrack [4]	0.223	0.732	0.161	0.838	0.056	0.981	0.147	0.850

Table B. **Quantitative results of affine-invariant depth on TAPVid-3D [9] minival split.** *Oracle tracker** rows use ground-truth 2D trajectories to determine the upper bound

Depth Estimator	Point Tracker	Aria [10] AbsRel ↓	Aria [10] δ_1 ↑	DriveTrack [1] AbsRel ↓	DriveTrack [1] δ_1 ↑	PStudio [6] AbsRel ↓	PStudio [6] δ_1 ↑	**Average** AbsRel ↓	**Average** δ_1 ↑
ZoeDepth [2]	Oracle tracker*	0.252	0.698	0.122	0.863	0.057	0.973	0.144	0.845
	CoTracker [8]	0.277	0.671	0.149	0.817	0.062	0.967	0.163	0.818
	LocoTrack [4]	0.282	0.676	0.175	0.799	0.063	0.966	0.173	0.814
Seurat (Ours)	CoTracker [8]	**0.244**	**0.701**	**0.139**	**0.845**	**0.060**	**0.967**	**0.148**	**0.838**
+ ZoeDepth [2]	LocoTrack [4]	0.266	0.694	0.178	0.818	0.063	0.964	0.169	0.825
DepthPro [2]	Oracle tracker*	0.180	0.742	0.148	0.827	0.067	0.970	0.132	0.846
	CoTracker [8]	0.219	0.708	0.186	0.778	0.074	0.958	0.160	0.815
	LocoTrack [4]	0.214	0.720	0.214	0.760	0.076	0.955	0.168	0.812
Seurat (Ours)	CoTracker [8]	**0.179**	0.757	**0.153**	**0.833**	**0.055**	**0.976**	**0.129**	**0.855**
+ DepthPro [3]	LocoTrack [4]	0.198	**0.760**	0.182	0.810	0.060	0.967	0.147	0.846

Table C. **Quantitative results on TAPVid-3D [9] using depth metrics with median scaling.** *Oracle tracker** rows use ground-truth 2D trajectories to determine the upper bound.

	Method	1 point	10 points	10^2 points	10^3 points	10^4 points
(I)	Tracking	2.07	2.02	2.02	4.58	34.98
(II)	Depth Inference	0.24	0.24	0.24	0.47	4.28
(III)	Overall	2.31	2.26	2.26	5.05	39.26

Table D. **Inference time and the number of query points.** We measure how the inference time (s) for a 24 frame video changes as varying the number of query points. We measure the time for point tracking (**I**) and depth inference with our method (**II**) separately. We use CoTracker [8] as a point tracker. Inference time is measured using Nvidia RTX 3090 GPU.

tial depth estimation in video, which we leave as future research. Furthermore, our use of a sliding window approach for processing long video sequences, while making depth variation manageable, somewhat limits the potential benefits of longer sequences. Exploring alternative approaches to effectively leverage extended temporal information is another interesting area for future research.

References

[1] Arjun Balasingam, Joseph Chandler, Chenning Li, Zhoutong Zhang, and Hari Balakrishnan. Drivetrack: A benchmark for long-range point tracking in real-world videos, 2023. A.1, A.2

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. A.1, A.2

[3] 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. *arXiv preprint arXiv:2410.02073*, 2024. A.1, A.2

[4] Seokju Cho, Jiahui Huang, Jisu Nam, Honggyu An, Seungryong Kim, and Joon-Young Lee. Local all-pair correspondence for point tracking. *arXiv preprint arXiv:2407.15420*, 2024. A.1, A.2

[5] 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. A.1, A.2

[6] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pages 3334–3342, 2015. A.1, A.2

[7] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. A.1

[8] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635v2*, 2023. A.1, A.2

[9] Skanda Koppula, Ignacio Rocco, Yi Yang, Joe Heyward, João Carreira, Andrew Zisserman, Gabriel Brostow, and Carl Doersch. Tapvid-3d: A benchmark for tracking any point in 3d. *arXiv preprint arXiv:2407.05921*, 2024. A.1, A.2

[10] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. A.1, A.2

[11] 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. A.1

[12] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. *arXiv preprint arXiv:2406.01493*, 2024. A.1, A.2

[13] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20417, 2024. A.1

[14] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. A.1