Appendix of Self-Supervised Monocular 4D Scene Reconstruction for Egocentric Videos

Chengbo Yuan[1,2], Geng Chen[2,3†], Li Yi[1,2], Yang Gao[1,2‡]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University
[2]Shanghai Qi Zhi Institute [3]UC San Diego
ycb24@mails.tsinghua.edu.cn gec001@ucsd.edu
{ericyi, gaoyangiiis}@mail.tsinghua.edu.cn

This appendix file provides:

For video visualization and online interactive demonstration, please refer to https://egomono4d.github.io/.

A. Details of Datasets

Figure 3 in the main paper visualizes samples from various datasets. Detailed information about the datasets is shown in Table 1. The data encompass different types of egocentric videos, with variations in camera motion (small/large), action complexity (simple/complex), and scene conditions (clean/cluttered).

For training, we use a combination of H2O[13] (40K frames), HOI4D[17] (540K frames), FPHA[9] (73K frames), EgoPAT3D[15] (823K frames), and Epic-Kitchen[4] (9.7M frames). The final training dataset contains a total of 11.2M frames, with Epic-Kitchen dominating the data (about 85%), providing large-scale diversity and a wide range of behavior modes. The other datasets contribute unique scene characteristics or behavior patterns. For instance, H2O[13] exclusively contains bi-manual operations with small camera motion in clean table scenes. Approximately 5% of the data is allocated for model validation.

For evaluation, we note a scarcity of egocentric datasets offering both high-quality depth and precise camera labels. We strongly encourage the computer vision and HOI communities to collect or synthesize larger-scale RGBD datasets with accurate pose annotations. We selected four datasets that meet the necessary criteria: H2O[13], HOI4D[17], ARCTIC[7], and POV-Surgery[35]. The sources for camera labels are provided in Table 1.

H2O[13] and HOI4D[17] feature simple scenes and actions and are used as test benchmarks for in-domain prediction performance. For zero-shot generalization evaluation, we use ARCTIC[7] and POV-Surgery[35]. ARCTIC[7] only provides labels for hands and objects, which is why we refer to it as ARCTIC-HOI, It is primarily used to assess the reconstruction quality of HOI. Both ARCTIC-HOI and POV-Surgery present significant challenges for EgoMono4D, as they exhibit a large domain gap from the training data. (1) For ARCTIC-HOI, the hand and object components occupy a much larger portion of the images compared to the training data. (2) For POV-Surgery, the dataset's unrealistic textures create a substantial visual domain gap, and the surgical scenes were not encountered during training.

B. Demonstration of UniDepth Backbone

Our architecture builds upon the UniDepth backbone [22], an encoder-decoder architecture depth estimator. UniDepth decouples the tasks of depth and camera estimation by transforming the scene from Cartesian coordinates to a pseudo-spherical representation, which enables dense camera prediction in spherical space. It then incorporates scene scale information from the camera prediction into the depth estimation module using Laplace Spherical Harmonic Encoding (SHE) and a cross-attention mechanism [5, 30]. For more details, please check out the original paper Piccinelli et al. [22].

† Work done during the internship at Shanghai Qi Zhi Institute. ‡ The corresponding author.

Training						
Datasets	**# of frames**	**Data Split**	**Camera Motion**	**Action**	**Scene**	**Note**
H2O	40K	Original Split	Small	Simple	Clean	Bi-manual
HOI4D	540K	Room ID	Medium	Simple	Clean	
FPHA	73K	Task	Medium	Complex	Clutter	
EgoPAT3D	823K	Scene ID	Large	Medium	Medium	Only pick & place
Epic-Kitchen	9.7M	Scene ID	Large	Complex	Clutter	
Evaluation						
Datasets	**# of frames**	**Label**	**Camera Motion**	**Action**	**Scene**	**Note**
H2O	8K	Calibration	Small	Simple	Clean	Bi-manual
HOI4D	12K	SfM	Medium	Simple	Clean	Contain noise
ARCTIC-HOI	13K	Mocap	Medium	Complex	Clean	Only hand and object label
POV-Surgery	26K	Synthesis	Large	Medium	Clutter	Unrealistic texture

Table 1. Comparison of Different Datasets for Training and Evaluation

C. Details of Evaluation Metrics

We provide the mathematical definitions of the metrics used to evaluate the performance of 3D point cloud sequence reconstruction and long-term 3D scene flow recovery.

C.1. Metrics for Pointclouds Sequence

We follow Örnek et al. [20] in using 3D Chamfer Distance (CD, measured in millimeters) and the 3D Pointclouds F-score (F, measured as a percentage %) to evaluate shape similarity. For implementation, we leverage the Kaolin library [8]. Given the ambiguity in the scale of pointclouds, we first align the predicted point cloud sequence to the ground truth using an estimated best-aligned global scaled SE(3) transformation (s, R, T).

3D Chamfer Distance (CD, mm). Given the predicted per-frame pointclouds $P \in R^{N\times 3}$ and the ground-truth $G \in R^{N\times 3}$, where N is the number of points, the 3D Chamfer Distance (CD) is defined as:

$$CD(R,G) = \sum_{x\in G} \min_{y\in R} ||x-y|| + \sum_{y\in R} \min_{x\in G} ||x-y|| \quad (1)$$

3D Pointclouds F-score (F). The 3D F-score combines precision and recall to provide a balanced evaluation of the predicted surface quality. In the context of 3D pointclouds, precision measures how many points from the predicted surface are close to the ground truth surface, while recall measures how many ground truth points are captured by the predicted surface. Given a distance δ as the positive threshold, the precision $P_\delta(R,G)$, recall $R_\delta(R,G)$ and F-score $F_\delta(R,G)$ are defined as:

$$P_\delta(R,G) = \frac{1}{|R|}\sum_{y\in R}[d_{y\to G} < \delta] \quad (2)$$

$$R_\delta(R,G) = \frac{1}{|G|}\sum_{x\in G}[d_{x\to R} < \delta] \quad (3)$$

$$F_\delta(R,G) = \frac{2P_\delta(R,G)R_\delta(R,G)}{P_\delta(R,G)+R_\delta(R,G)} \quad (4)$$

C.2. Metrics for Long-term 3D Scene Flow

Long-term 3D scene flow refers to predicting the future trajectories of multiple 3D query points in the pointclouds of the first frame [39]. Following previous works [2, 12, 39], we evaluate the precision of 3D flow recovery using three metrics: Average Displacement Error (ADE, measured in millimeters), Final Displacement Error (FDE, measured in millimeters), and Precision under Distance (P, measured as a percentage %). The 3D flow is generated by interpolating between the predicted and ground-truth pointclouds based on 2D tracking from CoTracker [10]. Before evaluation, we filter out trajectories affected by noise from flying pixels [23] using ground-truth depth information. To align the scale of scenes and the initial position of the 3D query points, we first perform a best-aligned scaled SE(3) transformation between the ground-truth and predicted pointclouds for the first frame.

Average Displacement Error (ADE, mm). Given the predicted 3D flow $F \in R^{T\times N\times 3}$ and ground-truth flow $G \in R^{T\times N\times 3}$, where T represents the number of frames and N represents the number of trajectories, ADE measures the average displacement across all timestamps.

	HOI4D				H2O				POV-Surgery†				ARCTIC-HOI†			
	ADE↓	FDE↓	P_5 ↑	P_{10} ↑	ADE↓	FDE↓	P_5 ↑	P_{10} ↑	ADE↓	FDE↓	P_5 ↑	P_{10} ↑	ADE↓	FDE↓	P_5 ↑	P_{10} ↑
Modularized Version	88.8	94.8	23.1	69	40.9	42.8	75.8	**97.0**	504.5	767.6	3.2	13.6	102.0	154.6	33.8	62.0
DS+UniDepth [29]	64.0	75.1	54.0	82.5	46.4	48.6	67.6	95.0	168.9	199.7	13.2	39.8	53.9	**72.1**	**60.1**	86.8
DUSt3R [36]	79.2	75.8	47.7	75.1	71.7	71.5	54.3	75.9	412.8	407.2	14.8	44.9	214.3	181.6	52.2	82.3
MonSt3R [40]	70.2	71.0	50.7	80.6	96.3	97.1	27.6	67.3	201.8	208.9	8.4	31.8	83.9	106.3	28.4	69.9
Align3R [18]	66.3	67.8	53.4	83.7	75.3	75.5	39.2	79.7	157.6	163.7	14.5	43.1	64.8	83.3	43.8	83.5
CUT3R [34]	72.5	72.6	49.8	81.3	50.7	57.4	67.9	91.3	304.6	353.0	4.6	19.3	84.3	106.2	27.4	69.9
EgoMono4D (Ours)	**57.3**	**60.6**	**59.2**	**87.0**	**37.0**	**38.9**	**77.0**	96.4	**125.0**	**138.9**	**19.1**	**55.2**	**53.8**	**72.1**	57.2	**87.3**

Table 2. The evaluation results for long-term 3D scene flow recovery are presented, with ADE (mm), FDE (mm), and Precision (P_δ, %). † denotes zero-shot generalization for EgoMono4D. For ARCTIC-HOI, the evaluation focuses solely on hand-object recovery quality. Overall, EgoMono4D significantly outperforms the other baselines.

(Only for HOI)	ADE↓	FDE↓	P_5↑	P_{10}↑
HOI4D	79.3 / **76.6**	84.6 / **78.4**	**48.3** / 46.2	77.8 / **81.0**
H2O	43.6 / **40.5**	45.8 / **41.4**	68.6 / **71.2**	95.7 / **98.2**
POV-Surgery†	196.0 / **192.2**	213.4 / **206.4**	**9.9** / **9.9**	32.4 / **32.8**

Table 3. Additional long-term 3D scene flow results on the HOI part. Values are presented **in the format 'DS+UniDepth / EgoMono4D (Ours)'**. Both models demonstrate comparable performance, with EgoMono4D showing a modest advantage.

$$ADE(F,G) = \frac{1}{N}\sum_{i=1}^{N}||F_i - G_i|| \tag{5}$$

Final Displacement Error (FDE, mm). FDE measures the displacement of the final timestamp.

$$FDE(F,G) = ||F_{T-1} - G_{T-1}|| \tag{6}$$

Precision under Distance (P, %). The precision metric measures the average percentage of points with an error within δ centimeters (cm).

$$P_\delta = \frac{1}{N}\sum_{i=1}^{N}[||F_i - G_i|| < \delta] \tag{7}$$

D. Long-term 3D Scene Flow Recovery Results

Task Long-term 3D scene flow [31] captures both the structure and dynamics of egocentric scenes in a compact format, making it valuable for various applications such as perception [25], autonomous driving [19], and robot learning [6, 24, 39]. Given a video and a set of query points in the first frame, the 3D flow [12, 39] represents the future trajectory of each query point in 3D space. Since egocentric datasets lack explicit 3D flow labels, we first employ CoTracker [10], a high-precision pixel tracker, to generate 2D long-term tracking (with a 35×35 grid of query points). These 2D trajectories are then unprojected using ground-truth pointclouds sequence to create 3D trajectory labels. We use the same method to obtain predictions for models.

Metric To align the scale of scenes , we first perform a best-aligned scaled SE(3) transformation between the ground-truth and predicted pointclouds for the first frame. We adopt metrics from related works [2, 12, 33, 39], including Average Displacement Error (ADE, mm), Final Displacement Error (FDE, mm), and Precision under Distance (P, %). We use the notation P_δ to represent precision with a δ centimeter (cm) threshold. Details are demonstrated in Appendix C.

Result Table 2 shows that our model outperforms all baselines for long-term 3D scene flow recovery on average. The visualizations are shown in Figure 1. DS+UniDepth performs well in hand-object motion estimation on the ARCTIC-HOI dataset. We also compared the HOI estimation of both models on three other datasets, with results in Table 3. Both models perform similarly, with EgoMono4D shows a modest advantage. However, DS+UniDepth's performance drops notably in full-scene estimation (e.g., POV-Surgery dataset) due to inconsistencies and accumulated errors between modules. Other models perform worse overall. Figure 1 illustrates the estimated long-term 3D scene flow. It can be seen that EgoMono4D successfully reconstructs 3D dynamics across diverse scenes.

E. Inference Speed Results

We evaluate the inference speed of different models (for 40 frames video). All measurements are conducted on a single NVIDIA GeForce 3090 GPU with a batch size of 1. Results are shown in Figure 6. Our model achieves 0.218 secs/frame speed, which is only slower than its modularized version [1] (0.215 secs/frame) and CUT3R [34] (0.115 sec/frame). However, the speed difference is very small, while our model achieves better reconstruction quality with large advantage.

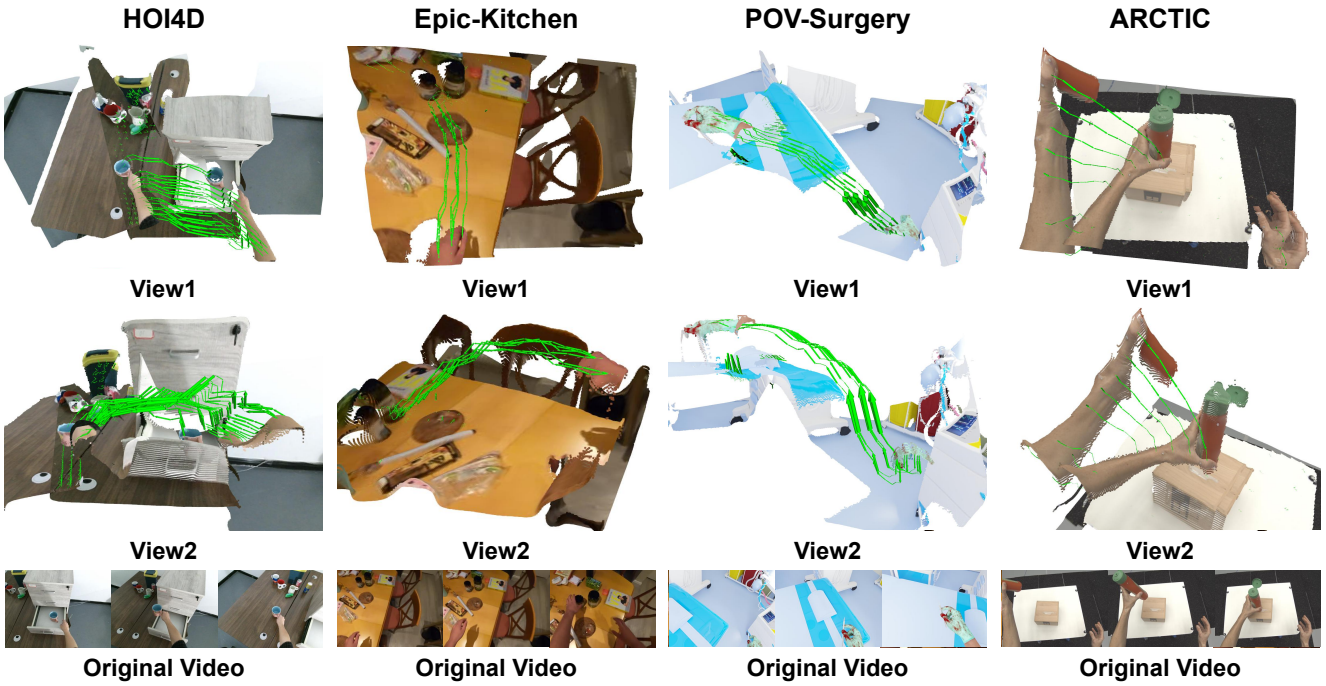

Figure 1. The visualization of the long-term 3D scene flow recovery. For clarity, we display only the pointclouds from the first and last frame. The green arrows representing the estimated 3D flow. EgoMono4D successfully recovers the motion of dynamic parts while maintaining other regions static to some extent.

N_w	N_o	HOI4D				POV-Surgery			
		CD↓	F_1 ↑	$F_{2.5}$ ↑	F_5 ↑	CD↓	F_1 ↑	$F_{2.5}$ ↑	F_5 ↑
4	**1**	5.9	27.9	59.6	83.1	33.8	13.5	32	53.9
2	**1**	17.9	11.5	29.4	51.6	/	/	/	/
8	**1**	6.4	26.0	55.6	80.0	35.2	12.6	30.8	53.1
12	**1**	6.7	25.1	53.5	78.4	/	/	/	/
4	**2**	5.9	27.9	59.7	83.1	34.6	13.4	32	53.9
4	**3**	5.9	27.9	59.7	83.1	35.0	13.4	31.9	53.8
8	**4**	/	/	/	/	36.3	16.6	30.5	52.6

Table 4. Comparison of pointclouds sequence reconstruction results across different window sizes (N_w) and overlapping sizes (N_o). Our model demonstrates robustness to variations in N_o, while maintaining consistency in N_w between training and inference is essential.

F. Ablation on Inference Strategy

During inference, our model processes N_w frames in a single feed-forward prediction. Theoretically, the window size N_w can be any value greater than 1. For videos with more frames than N_w, the overlapping size N_o between neighboring windows must also be determined. By default, we set $N_w = 4$ (consistent with training) and $N_o = 1$ (to optimize inference speed). We evaluate the impact of N_w and N_o on reconstruction performance using the pointclouds sequence reconstruction task on HOI4D [17] and POV-Surgery [35], with results shown in Table 4.

For the window size N_w, maintaining consistency between training and inference is crucial, likely because the video adapter is trained specifically to fuse 4 frames. Regarding overlapping size N_o, the model exhibits comparable performance for $N_o = 1, 2, 3$. Therefore, we select $N_o = 1$ to maximize inference speed.

G. Depth and Camera Results

We further evaluate the video depth and camera poses estimation performance on POV-Surgery [35], using the metrics from MonSt3R [40]. The results are presented in

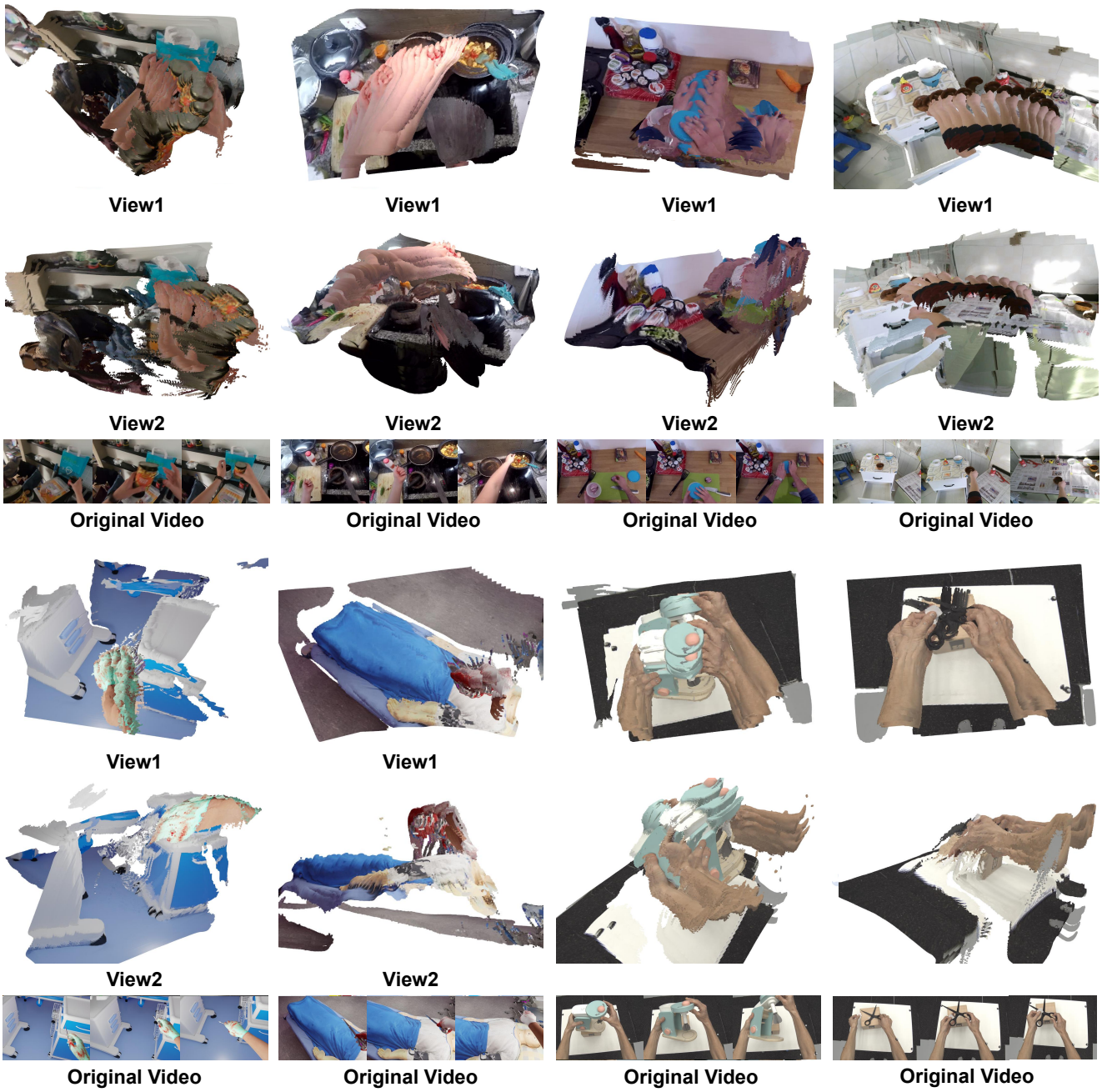

Figure 2. More visualization of pointclouds sequence reconstruction results from EgoMono4D.

Table 5. Our model does not outperform other methods in estimating these independent geometric variables. The depth and camera prediction performance of EgoMono4D is only comparable with other baseline methods. Instead, it enhances the consistency of them in 3D space, leading to improved pointclouds sequence reconstruction results. UniDepth [22] achieves the best depth estimation, while Align3R [18] provides the most accurate camera poses.

We believe this is due to two main reasons: (1) Our method is primarily optimized for 4D pointclouds reconstruction, rather than single geometry variable estimation. While it may not achieve the highest accuracy for individual variables, it ensures greater consistency and alignment among them. (2) Compared with supervised methods, self-supervised approach does not show a clear advantage when focusing solely on single-variable estimation, as indicated by previous work [26]. This is because there are more datasets available for single-variable estimation. We believe

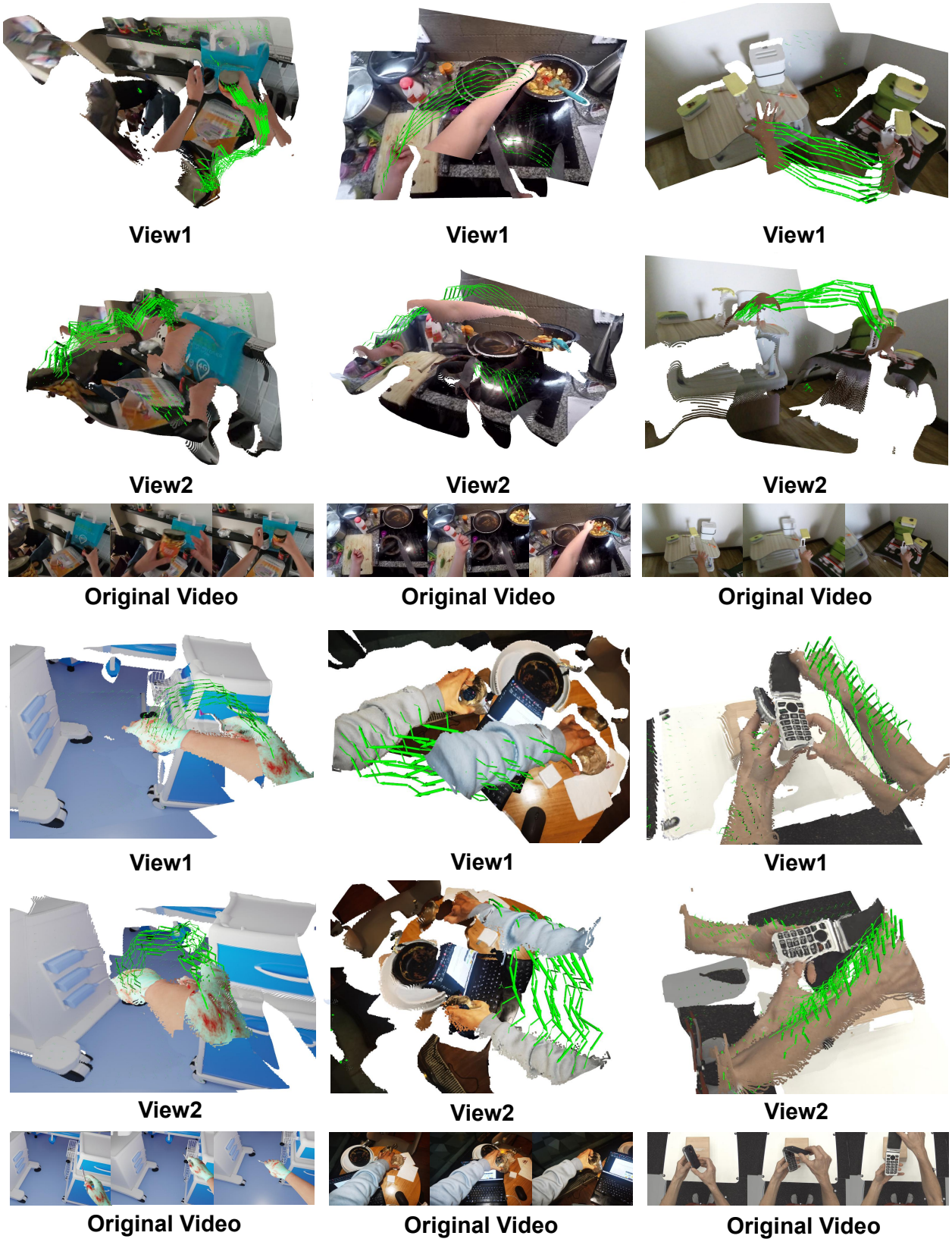

Figure 3. More visualization of long-term 3D scene flow recovery results from EgoMono4D.

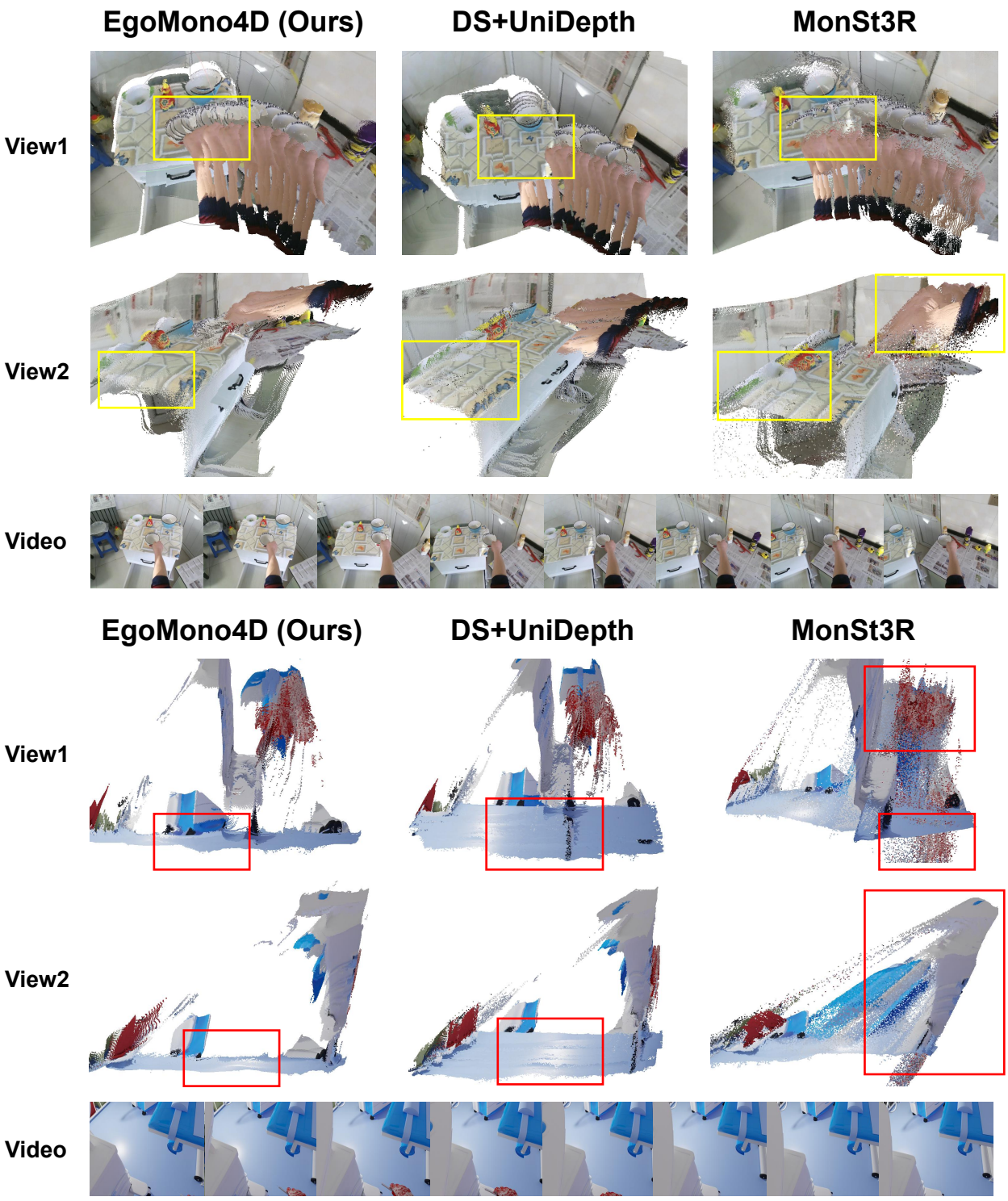

Figure 4. Qualitative comparison between EgoMono4D and other baseline methods. EgoMono4D demonstrates superior performance compared to baseline methods in both static scene reconstruction and dynamic HOI motion recovery. DS+UniDepth [22, 29] struggles with static scene alignment, while MonSt3R [40] exhibits limitations in accurately reconstructing hand geometry.

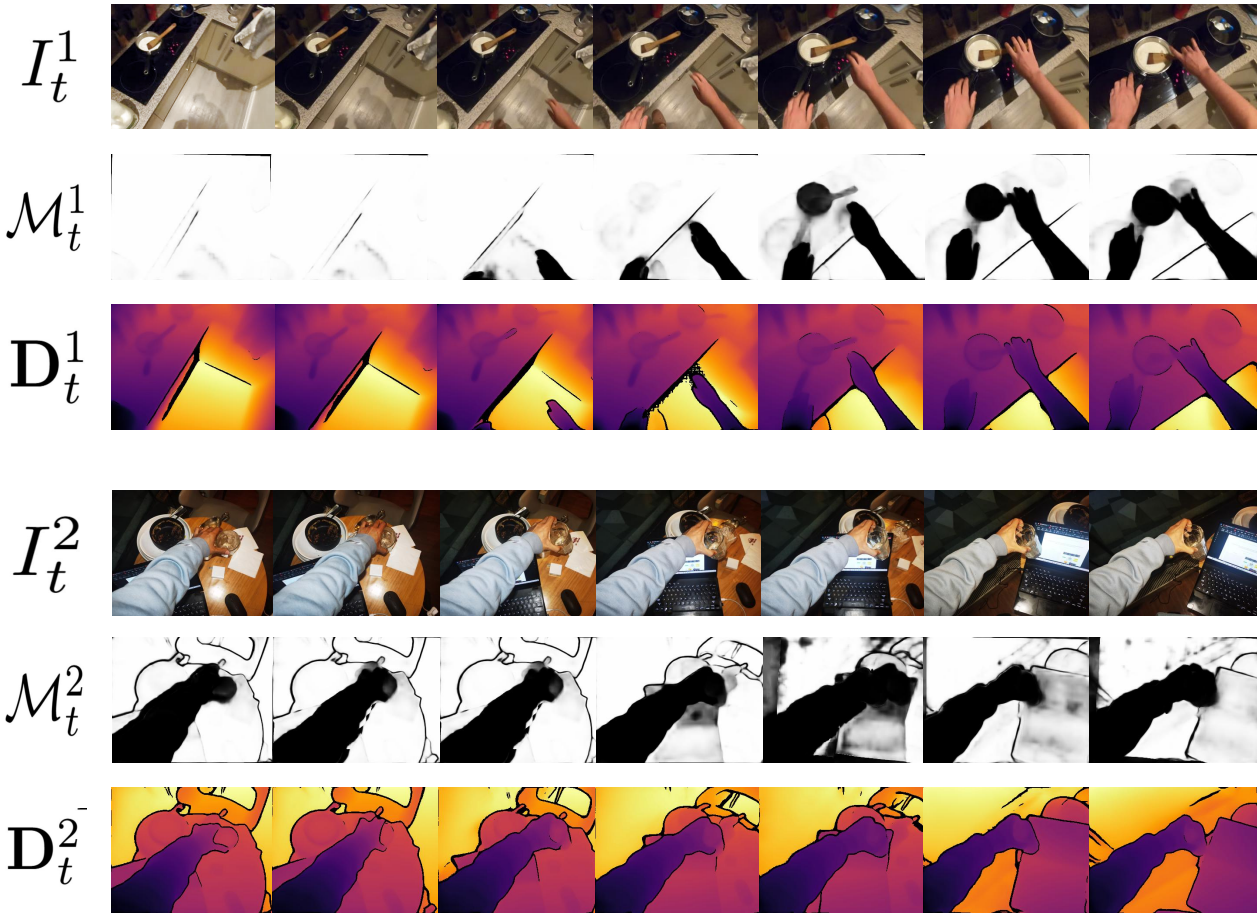

Figure 5. Visualization of original video I_t and the predicted procrustes-alignment confidence maps $\mathcal{M}_t$ and video depth D_t from EgoMono4D.

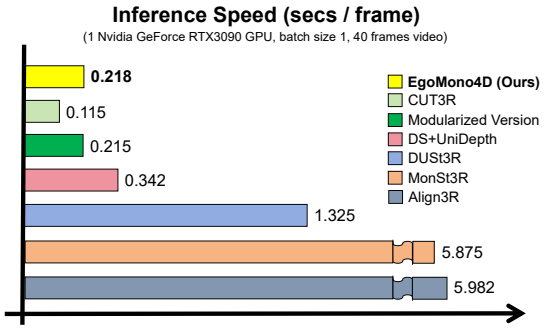

Figure 6. Inference speed comparison of different methods.

that using more advanced pretrained models and training on 0larger datasets may address this in the future [26, 27].

H. More Reconstruction Results Visualization

We provide more visualization of pointclouds sequence reconstruction results of EgoMono4D in Figure 2. More visualization of recovered long-term 3D scene flows are shown in Figure 3. Finally, we provide the visualization of intermediate geometric variables in Figure 5.

I. Qualitative Comparison with Baseline

We qualitatively compare EgoMono4D with other baseline methods on the dense point cloud sequence reconstruction task. Visual comparisons are presented in Figure 4. EgoMono4D demonstrates superior performance compared to baseline methods in both static scene reconstruction and dynamic HOI motion recovery. DS+UniDepth [22, 29] struggles with static scene alignment, while MonSt3R [40] exhibits limitations in accurately reconstructing hand geometry and dynamics.

J. Limitations and Future Directions

While EgoMono4D achieves impressive results in fast, dense, and generalizable dynamic HOI scene reconstruction, it still faces challenges with shape misalignment (see

	POV-Surgery (Video Depth)		
	AbsRel↓	$\delta_{0.05}$ ↑	$\delta_{0.1}$ ↑
UniDepth [22]	**11.9**	40.8	**64.7**
DUSt3R [36]	19.7	31.5	51.9
MonSt3R [40]	18.6	33.6	52.8
Align3R [18]	13.3	**41.7**	62.6
EgoMono4D (Ours)	<u>12.6</u>	<u>41.1</u>	<u>63.9</u>

	POV-Surgery (Camera Poses)		
	ATE↓	RPE-T↓	RPE-R↓
Modularized Version	47.03	4.10	4.18
DS+UniDepth [29]	9.05	4.17	0.39
MonSt3R [40]	<u>6.63</u>	<u>2.41</u>	<u>0.26</u>
Align3R [18]	**6.35**	**2.34**	**0.23**
CUT3R [34]	8.11	2.92	0.31
EgoMono4D (Ours)	11.54	4.01	0.43

Table 5. Result of POV-Surgery video depth and camera poses estimation. It shows that our model does not outperform others in estimating these geometric variables. Instead, our model focus on improving their consistency in 3D space, which leads to better pointclouds sequence reconstruction.

Figure 7. Two typical failure cases of EgoMono4D arise from inherent inconsistencies in UniDepth shape supervision.

Figure 5 in the main paper). This issue arises from inherent inconsistencies in UniDepth's [22] shape regularization and can be divided into two key problems. (1) Dynamic part (size) distortion: since we only regularize the shape of the dynamic part based on per-frame pointclouds predictions from UniDepth, the relative size of the dynamic part compared to the static part is determined by UniDepth's predictions. Any inaccuracy in this relative size may be carried over to EgoMono4D. (2) Static part misalignment: if the original predicted shapes of the static areas from UniDepth differ significantly between frames, it becomes difficult to adjust and align them consistently.

Although precise posed datasets are limited, there are more datasets available with depth labels. Training on these datasets with ground-truth shape supervision, or using a combination of labeled and unlabeled datasets, could help address these issues. Improvements in monocular depth estimation and intrinsic parameter estimation may also alleviate these problems.

Another limitation of EgoMono4D is that it currently supports video reconstruction only for videos with an FPS above a certain threshold. This is due to its reliance on the off-the-shelf optical flow estimation module [38], which may perform poorly with sparse views. Integrating correspondence, matching, or optical flow prediction directly into the model to enable fully end-to-end training could address this limitation [14, 32]. Sparse-view data also needs to be incorporated during training [36] to solve this problem.

Additionally, our model currently only supports a resolution of 288×384. Training models at higher resolutions could enable more flexible applications. Our model also fails when the dynamic portion of the image is too large or contains too many moving objects beyond the ones being manipulated. Although we focus on egocentric HOI scenes, our training paradigm could be extended to more general cases. We plan to train this general scene model using motion mask priors derived from motion segmentation [3, 37], salient video segmentation [28], and semantic segmentation [11, 16]. We also encourage the community to propose more synthetic datasets [21, 41] to explore supervised learning approaches [36, 40].

References

[1] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, pages 690–708. Springer, 2022. 3

[2] Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13702–13711, 2023. 2, 3

[3] Suhwan Cho, Minhyeok Lee, Seunghoon Lee, Chaewon Park, Donghyeong Kim, and Sangyoun Lee. Treating motion as option to reduce motion dependency in unsupervised video object segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5140–5149, 2023. 9

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 1

[5] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[6] Ben Eisner, Harry Zhang, and David Held. Flowbot3d: Learning 3d articulation flow to manipulate articulated objects. *arXiv preprint arXiv:2205.04382*, 2022. 3

[7] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 1

[8] Clement Fuji Tsang, Maria Shugrina, Jean Francois Lafleche, Towaki Takikawa, Jiehan Wang, Charles Loop, Wenzheng Chen, Krishna Murthy Jatavallabhula, Edward Smith, Artem Rozantsev, Or Perel, Tianchang Shen, Jun Gao, Sanja Fidler, Gavriel State, Jason Gorski, Tommy Xiang, Jianing Li, Michael Li, and Rev Lebaredian. Kaolin: A pytorch library for accelerating 3d deep learning research. https://github.com/NVIDIAGameWorks/kaolin, 2022. 2

[9] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 1

[10] 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. 2, 3

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 9

[12] 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. 2, 3

[13] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021. 1

[14] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 9

[15] Yiming Li, Ziang Cao, Andrew Liang, Benjamin Liang, Luoyao Chen, Hang Zhao, and Chen Feng. Egocentric prediction of action target in 3d. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20971–20980. IEEE, 2022. 1

[16] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 9

[17] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 1, 4

[18] Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. *arXiv preprint arXiv:2412.03079*, 2024. 3, 5, 9

[19] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 3

[20] Evin Pınar Örnek, Shristi Mudgal, Johanna Wald, Yida Wang, Nassir Navab, and Federico Tombari. From 2d to 3d: Re-thinking benchmarking of monocular depth prediction. *arXiv preprint arXiv:2203.08122*, 2022. 2

[21] 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. 9

[22] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 1, 5, 7, 8, 9

[23] Alexander Sabov and Jörg Krüger. Identification and correction of flying pixels in range camera data. In *Proceedings of the 24th Spring Conference on Computer Graphics*, pages 135–142, 2008. 2

[24] Daniel Seita, Yufei Wang, Sarthak J Shetty, Edward Yao Li, Zackory Erickson, and David Held. Toolflownet: Robotic manipulation with tools via predicting tool flow from point clouds. In *Conference on Robot Learning*, pages 1038–1049. PMLR, 2023. 3

[25] Lin Shao, Parth Shah, Vikranth Dwaracherla, and Jeannette Bohg. Motion-based object segmentation based on dense rgb-d scene flow. *IEEE Robotics and Automation Letters*, 3 (4):3797–3804, 2018. 3

[26] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Kick back & relax: Learning to reconstruct the world by watching slowtv. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15768–15779, 2023. 5, 8

[27] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Kick back & relax++: Scaling beyond ground-truth depth with slowtv & cribstv. *arXiv preprint arXiv:2403.01569*, 2024. 8

[28] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, Hanjing Su, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *IEEE Transactions on Multimedia*, 26:313–325, 2023. 9

[29] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 3, 7, 8, 9

[30] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[31] Sundar Vedula, Simon Baker, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 722–729. IEEE, 1999. 3

[32] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21686–21697, 2024. 9

[33] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024. 3

[34] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025. 3, 9

[35] Rui Wang, Sophokles Ktistakis, Siwei Zhang, Mirko Meboldt, and Quentin Lohmeyer. Pov-surgery: A dataset for egocentric hand and tool pose estimation during surgical activities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 440–450. Springer, 2023. 1, 4

[36] 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*, pages 20697–20709, 2024. 3, 9

[37] Junyu Xie, Charig Yang, Weidi Xie, and Andrew Zisserman. Moving object segmentation: All you need is sam (and flow). *arXiv preprint arXiv:2404.12389*, 2024. 9

[38] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022. 9

[39] Chengbo Yuan, Chuan Wen, Tong Zhang, and Yang Gao. General flow as foundation affordance for scalable robot learning. *arXiv preprint arXiv:2401.11439*, 2024. 2, 3

[40] 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. 3, 4, 7, 8, 9

[41] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19855–19865, 2023. 9