4D Primitive-Mâché: Glueing Primitives for Persistent 4D Scene Reconstruction

Supplementary Material

1. System Analysis

We analyse the behaviour of our system across several dimensions.

Robustness to geometry frontend Since our method relies on the outputs of feed-forward geometry estimation networks, we first examine its sensitivity to the choice of model. Table Sec. 1 shows that 4DPM is robust to the choice of frontend, yielding a consistent improvement in F-score across all configurations (by at least $+0.235$).

Component ablation Table Sec. 1 presents a detailed ablation of our system's components. Our correspondence filtering mechanism is critical to performance: removing it causes a drop in F-score of 0.15 (-19%). Second-order optimisation substantially outperforms first-order Adam. Replacing Alltracker with SEA-RAFT [4] results in a modest performance degradation. Finally, our method is robust to segmentation quality: using unfiltered SAMv2 masks or the lighter SAMv2-Small variant yields performance comparable to the full pipeline.

	F-score	delta F-score
π^3	0.5219	-
VGGT	0.5330	-
DAv3	0.4587	-
π^3 + **Ours**	0.7573	**+0.235**
VGGT + **Ours**	0.7700	**+0.237**
DAv3 + **Ours**	0.7103	**+0.256**

Table 1. **Robustness to different geometry estimators on HO3D.**

	F-score
Unfiltered correspondence	0.6401
FO optimiser	0.7228
SEA-RAFT for correspondence	0.7727
Unfiltered masks from SAMv2	0.7765
SAMv2-small	0.7947
Ours	**0.7948**

Table 2. **Ablation study on Multi-obj dataset.**

2. Additonal Qualitative Results

3. Analytical Jacobian Derivation

We provide analytical Jacobians for a single pairwise residual of an object. In practice, the Hessian is block-diagonal with respect to objects.

Given relative residuals $r = T_j^{-1} T_i X_i - \widehat{X_j}$, we derive its analytical right Jacobians $\mathbf{J} = [\mathbf{J}_{T_i} | \mathbf{J}_{T_j}]$ with respect to object poses T_i and T_j.

Let $Z = T_j^{-1} T_i$, then:

$$\frac{dZ}{dT_i} = \mathrm{Id} \tag{1}$$

$$\frac{dZ}{dT_j} = -\operatorname{Ad}\left(Z^{-1}\right) \tag{2}$$

From [3], if $Z = \begin{bmatrix} R_Z & t_Z \\ 0 & 1 \end{bmatrix}$, $R_Z \in \mathbb{SO}(3)\,, t_Z \in \mathbb{R}^3$ and Act is the action operator of $\mathbb{SE}(3)$ group on $\mathbb{R}^3$, i.e $\operatorname{Act}(Z, X) = R_Z X + t_Z$, then:

$$\frac{d\operatorname{Act}(Z,X)}{dZ} = \begin{bmatrix} R_Z & -R_Z [X]_\times \end{bmatrix} \tag{3}$$

Then, using chainrule,

$$\begin{aligned} \frac{dr}{dT_i} &= \frac{d(ZX_i - \widehat{X_j})}{dT_i} = \\ &= \frac{d\operatorname{Act}(Z,X)}{dZ} \frac{dZ}{dT_i} = \begin{bmatrix} R_Z & -R_Z [X_i]_\times \end{bmatrix} \end{aligned} \tag{4}$$

$$\begin{aligned} \frac{dr}{dT_j} = \frac{d(ZX_i - \widehat{X_j})}{dT_j} = \frac{d\operatorname{Act}(Z,X_i)}{dZ} \frac{dZ}{dT_j} = \\ = \begin{bmatrix} R_Z & -R_Z [X_i]_\times \end{bmatrix} \left(-\operatorname{Ad}\left(Z^{-1}\right)\right) \end{aligned} \tag{5}$$

The last equation can be simplified further. Recall that $Z^{-1} = \begin{bmatrix} R_Z^T & -R_Z^T t_Z \\ 0 & 1 \end{bmatrix}$ and, therefore, its adjoint is:

$$\operatorname{Ad}(Z^{-1}) = \begin{bmatrix} R_Z^T & [-R_Z t_Z]_\times R_Z^T \\ 0 & R_Z^T \end{bmatrix} \tag{6}$$

Then the second-block column of $\frac{dr}{dT_j}$ corresponding to rotation is:

$$\begin{aligned} &- R_Z [-R_Z t_Z]_\times R_Z^T + R_Z [X_i]_\times R_Z^T = \\ &= [t_Z]_\times + [R_Z X_i]_\times = \\ &= [R_Z X_i + t_Z]_\times = [\operatorname{Act}(Z, X_i)]_\times \end{aligned} \tag{7}$$

4. First Order vs Second Order Study

We compare the performance of our system using Gauss-Newton optimisation against Adam. In Sec. 4, we report the resulting F-score on our Multi-Object dataset for our

CVPR
#41171

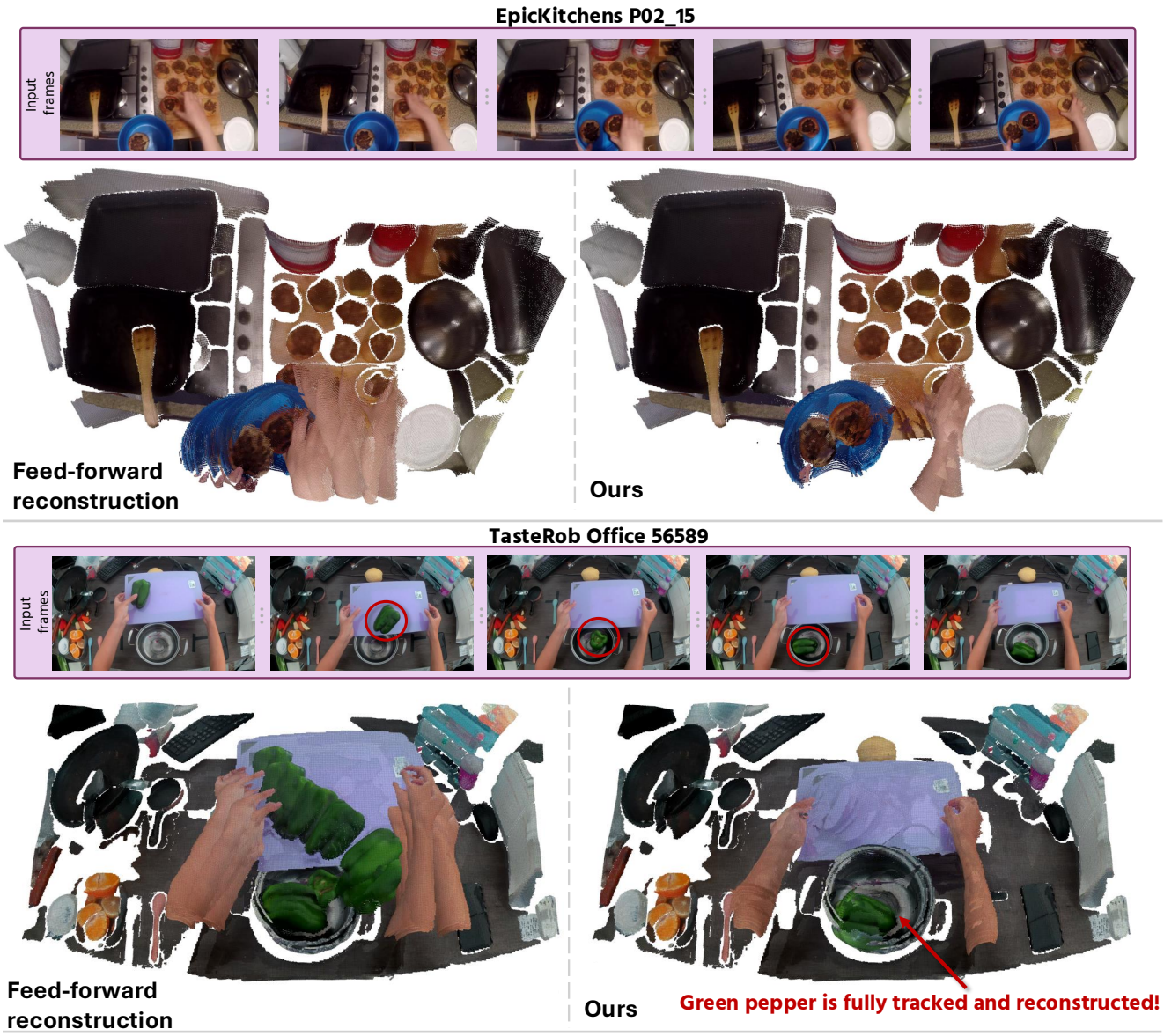

Figure 1. **Reconstruction on ego-centric data.** Our system runs of the box on different data distributions, such as ego-centric data, including TasteRob [5] and EpicKitchens [1] datasets.

system compared to a variant using the first-order Adam optimiser [2]. Interestingly, even with a generous time budget, the first-order optimiser never converged to the same level of quality, both quantitatively and qualitatively. In our experiments, the translational component of the transformation was often estimated correctly, while the rotational component remained problematic.

	F-Score	Time Spent
Adam 500 steps	0.6342	20s
Adam 1k steps	0.6474	40s
Adam 10k steps	0.7228	400s
Ours (10 steps)	0.7843	2s
Ours (50 steps)	**0.7948**	**10s**

Table 3. **Reconstruction quality of our method with different optimisers and computational budgets.**

5. Run-Time Analysis

In Sec. 5, we report a performance breakdown of our system. The frontend dominates overall runtime, with video segmentation being the most expensive component as we propagate masks for all potential objects in the scene (typi-

CVPR
#41171

Figure 2. **Failure Cases.** Our system struggles with rapid motion (e.g. shaking the milk carton) and object re-identification when hands disappear from and re-enter the camera view.

Component	Time Spent	Peak Memory
Frontend (correspondence)	2.3s	6.9GB
Frontend (π^3)	5.6s	8.7GB
Frontend (segmentation)	42s	13.1GB
Frontend (combined)	50s	-
Backend	9.6s	14.5GB
Motion Segmentation	2s	7.6GB

Table 4. **Performance and Memory benchmarking.**

cally around 50-100 objects). Backend performance scales with image resolution: our dense alignment at 512×512 takes approximately 10 seconds, reducing to 2.5 seconds at 256×256.

Further optimisation of the backend is possible, though the frontend now represents the primary opportunity for improvement. Additionally, our current implementation does not include early termination of the optimisation; incorporating this could yield further performance gains.

References

[1] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 2

[3] Joan Sola, Jeremie Deray, and Dinesh Atchuthan. A micro lie theory for state estimation in robotics. *arXiv preprint arXiv:1812.01537*, 2018. 1

[4] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1

[5] Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, and Xiaoguang Han. Taste-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2