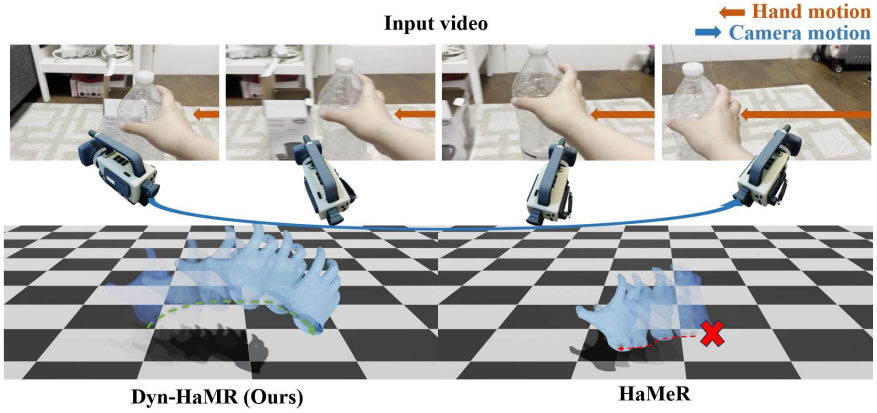
IMPERIAL

Dyn-HaMR: Recovering 4D Interacting Hand Motion from a Dynamic Camera

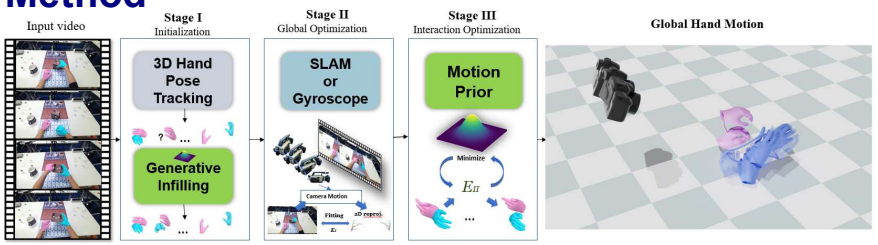
Zhengdi Yu (z.yu23@imperial.ac.uk)
Stefanos Zafeiriou (s.zafeiriou@Imperial.ac.uk)
Tolga Birdal (tbirdal@imperial.ac.uk)

Motivation



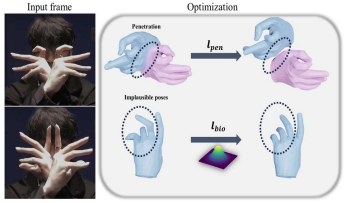
Dyn-HaMR (Ours) can disentangle the camera and object poses to recover the 4D global hand motion in the real world whilst state-of-the-art 3D hand reconstruction methods like HaMeR [1] fail to do so.

Method



Our main contributions include:

- Introducing the first optimization-based approach capable of disentangling and reconstructing global 4D pose and shape of two hands, and camera trajectory.
- Proposing a data-driven hand motion prior combined with biomechanical constraints, allowing realistic and complex hand interactions to guide the optimization.
- Conducting comprehensive experiments on challenging in-the-wild videos and benchmarks, demonstrating substantial performance improvements over state-of-the-art methods in 4D global motion recovery.



$$E_{II}({}^{w}q^{h}, \omega, \mathbf{R}_t, \tau^{c}_t) = \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{pen}} + \mathcal{L}_{\text{bio}} + \lambda_{2d}\mathcal{L}_{2d} + \lambda_s\mathcal{L}_{\text{smooth}} + \lambda_{\text{cam}}\mathcal{L}_{\text{cam}} + \lambda_J\mathcal{L}_J + \lambda_\beta\mathcal{L}_\beta.$$

Qualitative Results

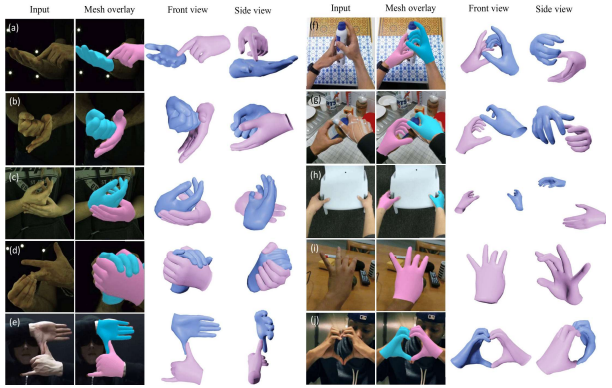


Fig 3. Qualitative evaluation on InterHand2.6M [5]. In each row, we show the mesh overlay and detailed reconstruction from different views.

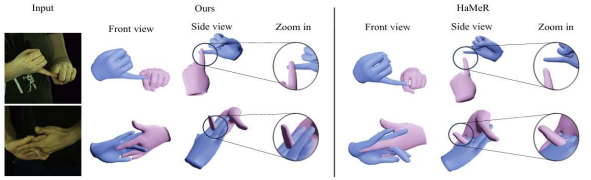


Fig 4. Comparison with state-of-the-art hand reconstruction approach [1] (static camera) on InterHand2.6M dataset [5].

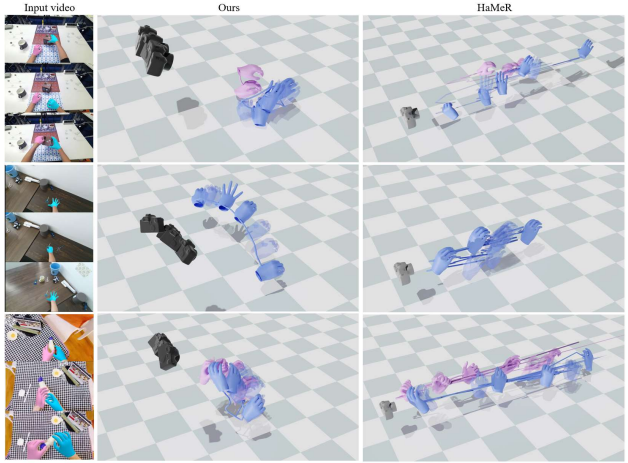


Fig5. Qualitative comparison with state-of-the-art method HaMeR [1]. It can be seen that our method recovers significantly more plausible global hand motion. first row is from H2O dataset [19], while second & third rows are from HOI4D dataset [28].

Quantitative Results

Method	MPJPE ↓	MPVPE ↓	Acc Err ↓
InterWild [30]	12.35	13.45	6.68
DIR [40]	9.09	9.43	8.92
ACR [52]	8.75	9.01	3.99
IntagHand [23]	9.26	9.71	4.41
HaMeR [36]	9.84	10.13	5.13
Ours (w/o III)	8.98	9.25	4.72
Ours (Dyn-HaMR)	**7.94**	**8.15**	**2.76**

Tab 1. Quantitative evaluation results for InterHand2.6M [5] 30 fps dataset. We compare our method with the state-of-the-art hand reconstruction methods on local hand poses.

Method	G-MPJPE ↓	GA-MPJPE ↓	MPJPE ↓	Acc Err ↓
ACR [52]	113.6	88.5	46.8	14.3
IntagHand [23]	105.5	81.5	45.6	13.5
HaMeR [36]	96.9	75.7	32.9	9.21
Ours (w/o III)	51.9	41.2	24.9	9.5
Ours (Dyn-HaMR)	**45.6**	**34.2**	**22.5**	**4.2**

Tab 2. Quantitative evaluation results for H2O [4] dataset. Our method demonstrates significant improvements over state-of-the-art approaches in recovering both local and global 4D hand motion, with additional gains achieved when incorporating Stage III.

Method	H2O				InterHand2.6M			
	Jerk ↓	Pen ↓	Trans Err ↓	FID ↓	Jerk ↓	Pen ↓	Trans Err ↓	FID ↓
ACR [52]	149.43	0.07	10.89	1.95 / 4.45	153.62	5.05	8.65	2.51 / 5.36
IntagHand [23]	166.38	0.06	11.15	2.14 / 4.12	165.31	4.82	9.19	2.69 / 5.07
HaMeR [58]	195.77	0.06	10.43	1.76 / 4.78	183.45	5.17	8.43	2.45 / 5.45
Ours (w/o bio. const.)	2.65	0.04	4.71	1.89 / 2.78	4.57	2.67	4.41	1.89 / 4.12
Ours (w/o pen. const.)	2.36	0.02	4.13	1.38 / 2.12	**4.03**	4.23	4.93	1.53 / 4.64
Ours (w/o III)	2.98	0.02	4.21	2.01 / 2.93	4.81	4.49	4.96	2.89 / 4.87
Ours (Dyn-HaMR)	**2.34**	0.009	**5.67**	**1.34 / 1.98**	4.26	**2.46**	**4.35**	**1.49 / 3.56**

Tab3. Plausibility evaluation on multiple datasets. Results are reported on the H2O [4] and InterHand2.6M [5] to analyze the jitter, penetration, translation, and plausibility. FID is reported for both single hand (left) and two hands (right).

Method	G-MPJPE ↓	GA-MPJPE ↓	MPJPE ↓	Acc Err ↓
Stage I	84.5	72.5	25.6	8.8
Stage I+II	51.9	41.2	24.9	9.5
w/o bio. const.	49.6	43.1	24.5	4.3
w/o pen. const.	46.3	34.7	23.6	**4.1**
w/o gen. infill.	48.9	37.8	24.1	5.6
Ours (Dyn-HaMR)	**45.6**	**34.2**	**22.5**	4.2

Tab 4. Ablation of pipeline components on H2O [4] dataset. It shows the impact of removing different components from the pipeline on various performance metrics.

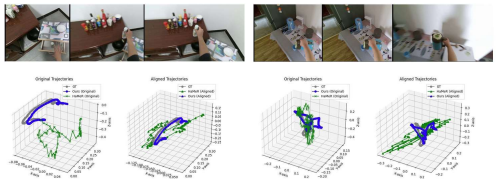


Fig 4. Comparison of global trajectory on HOI4D [6].

References

[1] G. P et al., Recon structing hands in 3d with transformers. **CVPR2024**
[2] Ye, V., el al., Decoupling human and camera motion from videos in the wild. **CVPR 2023**
[3] Duran, E., et al., Hmp: Hand motion priors for pose and shape estimation from video. **WACV 2024**
[4] T. Kwon et al., H2o: Two hands manipulating objects for first person interaction recognition. **ICCV 2021**
[5] G. Moon et al., Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb im age.. **ECCV 2020**
[6] Y. Liu et al., Hoi4d: A 4d egocentric dataset for category-level human object interaction.. **CVPR 2022**