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

Zhengdi Yu[1], Alara Dirik[1], Stefanos Zafeiriou[1], Tolga Birdal[1]
[1] Imperial College London
{z.yu23, a.dirik22, s.zafeiriou, t.birdal}@imperial.ac.uk

Abstract

Existing monocular hand reconstruction methods typically adopt the weak perspective camera model to simulate the hand motion in the camera frustum. As a result, they fail to recover 3D global trajectory and yield incorrect depth reasoning when the input is recorded by dynamic cameras, especially in egocentric hand interaction videos. In this paper, we make the first attempt to reconstruct 4D global hand motion from monocular videos recorded with dynamic cameras in the wild. Furthermore, To model accurate interacting 3D hands, we leverage a learned data-driven hand motion prior to explicitly refine the plausible and complex 3D interaction. Our method significantly outperforms the state-of-the-art approaches both qualitatively and quantitatively on challenging datasets with dynamic cameras.

1. Introduction

Existing monocular hand reconstruction methods [1, 2, 6, 8, 9, 12, 13, 21, 23–25] typically adopt the weak perspective camera model to simulate the global hand motion. As shown in Fig. 1, existing hand reconstruction works estimate hand poses in the camera coordinate system frustum or even root-relative coordinate system [12, 13], while the interaction with objects in the 3D virtual world will require estimating 4D hand motions in global coordinates consistent with the accurate localization of the objects in the scene. In contrast to all the existing works, we present an optimization-based pipeline **Dyn-HaMR**, which aims to reconstruct 4D global two hands from complex scenes with even dynamic cameras while modeling better interaction with a learned hand motion prior through an optimization routine, without the need for a sensor, multi-view camera setup, or any prior knowledge. To fully showcase the superiority of our method, we conduct extensive experiments on the captured in-the-wild dynamic interaction videos as well as existing benchmarks. Our method outperforms existing state-of-the-art approaches significantly in terms of the 4D global motion recovery.

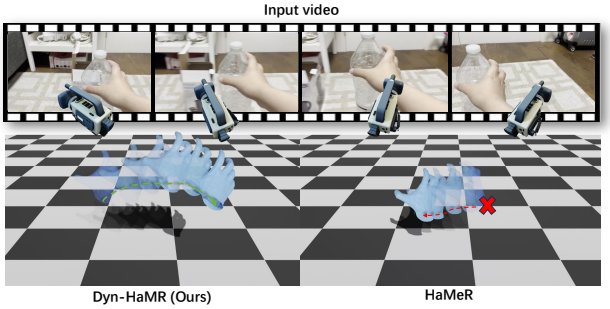

Figure 1. **Dynamic camera problem.** The **green** and **red** arrows represent the direction of the hand motion. Dyn-HaMR(Ours) can recover the 4D global hand motion in the real world whilst state-of-the-art 3D hand reconstruction methods [9, 14, 21] fail to do so.

2. Method

We consider an input video $\mathcal{V} = \{\mathbf{I}_1, \cdots, \mathbf{I}_T\}$ with T frames containing two, possibly interacting hands undergoing arbitrary 6D camera motion. We aim to recover the global trajectory of the hands in the *world coordinate system*. As illustrated in Fig. 2, our three-stage pipeline, inspired by recent dynamic human motion perception works [16, 20], includes: (I) Using state-of-the-art hand pose estimation [14, 19, 21] and motion priors [3] to initialize hand states per frame in the camera coordinates. (II) Applying an advanced SLAM system for initial global motion. (III) Refining hand displacements, interactions, penetration, and biomechanical constraints using learned motion priors.

Representing hand motion At time t, we parameterize hand pose and shape using the MANO model [15] as $\boldsymbol{q}_t^h = \boldsymbol{\theta}_t^h, \boldsymbol{\beta}_t^h, \boldsymbol{\phi}_t^h, \tau_t^h$, where $\boldsymbol{\theta}_t^h \in \mathbb{R}^{3\times 15}$ is the local hand pose with 15 joints, $\boldsymbol{\beta}_t^h \in \mathbb{R}^{10}$ are shape coefficients, and $(\boldsymbol{\phi}_t^h, \tau_t^h)$ denote the global wrist pose (root orientation $\boldsymbol{\phi}_t^h \in \mathbb{R}^3$ and translation $\tau_t^h \in \mathbb{R}^3$). The handedness is denoted by $h \in [l, r]$. We assume $\boldsymbol{\beta}^h$ remains constant throughout the sequence. We use ${}^{\mathrm{c}}\boldsymbol{q}_t^h$ and ${}^{\mathrm{w}}\boldsymbol{q}_t^h$ for camera and world coordinates, respectively, and ${}^{\mathrm{c}}\mathbf{Q}^h$ and ${}^{\mathrm{w}}\mathbf{Q}^h$ for trajectories. The MANO parameters $\boldsymbol{q}_t^h$ recover hand mesh vertices $\mathbf{V} \in \mathbb{R}^{3\times 778}$ and joints $\mathbf{J} \in \mathbb{R}^{3\times 21}$ through differentiable functions $M(\boldsymbol{J}_t^h, \boldsymbol{\beta}^h, \boldsymbol{\phi}_t^h, \tau_t^h)$.

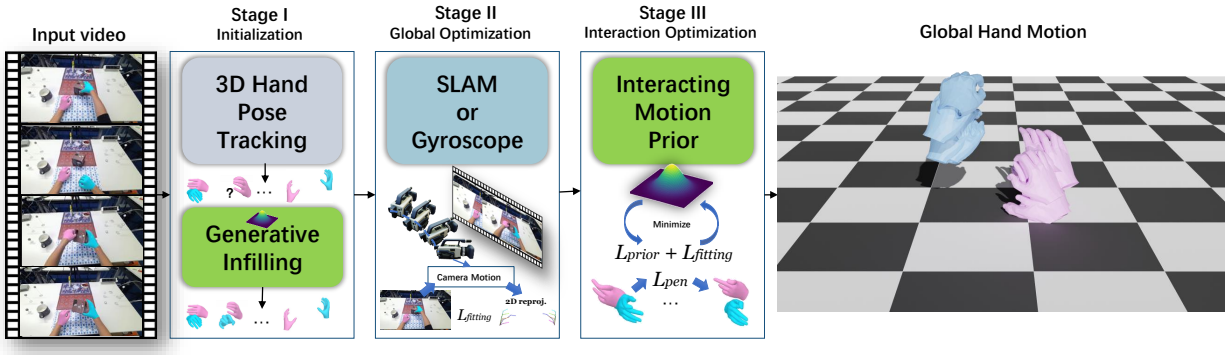

Figure 2. Overview of our method. We design a three-stage optimization pipeline to recover the 4D global hand motion from in-the-wild videos even recorded with dynamic cameras.

2.1. Stage I: Hierarchical Initialization

In stage I, we initialize the motion state $\boldsymbol{q}_t^h$ by the efficient two-hand tracking system with a hierarchical fusion scheme, which is followed by the generative infilling.

Motion infilling and temporal consistency Typical single-frame interacting hand reconstruction methods naturally lack temporal coherence and there could be missed detection due to the frequent occlusion during interactions, making the trajectory ${}^{c}Q^h$ *incomplete*. We address both problems by employing the hand motion prior [3] as a generative, smooth motion hallucinator. To do so, for both hands, we optimize for the latent code $\mathbf{z}^h$ in HMP so as to fit the frames where detections are present. We initialize this optimizer by using a canonical *slerp* interpolation in the pose space, finally leading to the initial 4D hand trajectory in the camera coordinate system ${}^{c}\mathbf{Q}^h$.

Initialization of 2D observations we incorporate ViTPose [19] and MediaPipe [11] with the reprojection of [14] and subsequently feed through a confidence-guided filter. To fill in the missed 2D keypoint detections, we reproject the 3D keypoints ${}^{c}\mathbf{J}_t^h$ from ${}^{c}\boldsymbol{q}_t^h$ onto 2D by weak-perspective projection as $\hat{\mathbf{J}}_\mathbf{t}^\mathbf{h} \in \mathbb{R}^{3\times 21}$.

2.2. Stage II: 4D Global Motion Optimization

Given the trajectory in the camera coordinate system ${}^{c}\mathbf{Q}$ (by Sec. 2.1), our key idea here is to compute the relative camera motion with a data-driven SLAM system, DPVO [18], and to estimate the transform $\mathbf{C}_t = \{\mathbf{R}_t, \boldsymbol{T}_t\}$ at each timestep t from the camera coordinate system to the world coordinate system. Then, the composition of hand motion ${}^{c}\mathbf{Q}^h$ and camera motion, *i.e.*, ${}^{w}\boldsymbol{q}_t^h = \mathbf{C}_t \odot {}^{c}\boldsymbol{q}_t^h$ reveals the global motion. We also optimize a world scale factor ω to explicitly model the relative scale between the displacements of the camera and hand motion inspired by [20].

Optimization variables During optimization, we take as input the initialized 2D keypoints sequence $\hat{\boldsymbol{J}}_t^h$, the 3D motion state sequence ${}^{c}\mathbf{Q}^h$ in the camera coordinate system, and the world-to-camera transformation $\mathbf{C}_t$ estimated by the SLAM system, and subsequently propose to jointly optimizes the global trajectories, orientation, and local hand poses and the camera extrinsics $\mathbf{C}_t$ to match the 2D observations. Specifically, we initialize ${}^{w}\mathbf{Q}$ as follows:

$$ {}^{w}\boldsymbol{\phi}_t^h = \mathbf{R}_t^{-1} \cdot {}^{c}\boldsymbol{\phi}_t^h \quad \text{and} \quad {}^{w}\tau_t^h = \mathbf{R}_t^{-1} \cdot {}^{c}\tau_t^h - \omega \mathbf{R}_t^{-1} \cdot \boldsymbol{T}_t, \tag{1} $$

while ${}^{w}\boldsymbol{\theta}_t^h = {}^{c}\boldsymbol{\theta}_t^h = \boldsymbol{\theta}_t^h$ and ${}^{w}\boldsymbol{\beta}_t^h = {}^{c}\boldsymbol{\beta}_t^h = \boldsymbol{\beta}^h$ remain the same. The initial world-camera scale factor is set as $\omega = 1$. With the initialized motion state ${}^{w}\boldsymbol{q}_t^h$, the 3D mesh joints at each timestep, ${}^{w}\mathbf{J}_t^h$, can be extracted as:

$$ {}^{w}\mathbf{J}_t^h = \mathbf{L} \cdot W(H(\boldsymbol{J}_t^h, \boldsymbol{\beta}^h), P(\boldsymbol{\beta}^h), \mathbf{S}) + {}^{w}\tau_t^h \tag{2} $$

Optimization scheme We optimize the trajectory in the world frame by minimizing the following objectives:

$$ \begin{aligned} E({}^{w}\boldsymbol{q}^h, \omega, \mathbf{R}_t, \boldsymbol{T}_t) = \lambda_{2d}\mathcal{L}_{2d} + \lambda_s\mathcal{L}_{\text{smooth}} + \lambda_{\text{cam}}\mathcal{L}_{\text{cam}} \\ + \lambda_{\boldsymbol{J}}\mathcal{L}_{\boldsymbol{J}} + \lambda_{\boldsymbol{\beta}}\mathcal{L}_{\boldsymbol{\beta}}. \end{aligned} \tag{3} $$

The first term aligns ${}^{w}\mathbf{Q}^h$ with the 2D observations:

$$ \mathcal{L}_{2d} = \sum_{t=0}^{T} \sum_{h\in\{l,r\}} \rho\left(\mathbf{C}_t^h \left(\tilde{\boldsymbol{J}}_t^h - \hat{\boldsymbol{J}}_t^h\right)\right) \tag{4} $$

Here, $\tilde{\boldsymbol{J}}_t^h = \Pi({}^{w}\mathbf{J}_t^h, \mathbf{R}_t, \omega, \boldsymbol{T}_t, \mathbf{K})$ is the 2D reprojection of current 3D keypoints and Π is the perspective camera projection with intrinsic K. $\mathbf{C}_t^h$ is a mask of the joint visibility $\rho(\cdot)$ is the Geman-McClure robust function [5]. Essentially, the motions of the two hands will further constrain the camera scale factor ω and improve the performance of complex hand interaction videos. To further eliminate implausible

poses, we leverage the temporal information and introduce regularization terms:

$$\mathcal{L}_{\text{smooth}} = \sum_{t=0}^{T} \sum_{h\in\{l,r\}} \|{}^{\text{w}}\mathbf{J}_{t+1}^{h} - {}^{\text{w}}\mathbf{J}_{t}^{h}\|^2 + \sum_{t=0}^{T} \sum_{h\in\{l,r\}} d_{\boldsymbol{\theta}} \left(\boldsymbol{\theta}_{t+1}^{h} \ominus \boldsymbol{\theta}_{t}^{h}\right)^2, \quad (5)$$

$$\mathcal{L}_{\text{cam}} = \sum_{t=0}^{T} d_{\mathbf{R}}(\mathbf{R}_{t+1}, \mathbf{R}_t)^2 + \sum_{t=0}^{T} \|\boldsymbol{T}_{t+1} - \boldsymbol{T}_t\|^2. \quad (6)$$

where rotational terms are geodesic distances. For reasonable regularization terms, to reduce the jittery poses, we employ standard pose $\mathcal{L}_{\boldsymbol{J}} = \sum_{t=0}^{T}\sum_{h\in\{l,r\}} \|\boldsymbol{J}_t^h\|^2$ and shape prior $\mathcal{L}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^h) = \sum_{h\in\{l,r\}} \|\boldsymbol{\beta}^h\|^2$ term [15].

2.3. Stage III: Motion Prior Optimization

After obtaining global hand motion in the world coordinate system from previous stages. In Stage III, we introduce an interacting hand motion prior optimization module to better model the interactions with plausibility, which also helps to determine the contribution of camera motion from the hand motion with a well-learned scale factor ω.

Optimization variables For the latent optimization with motion prior, we omit the decoded global orientation as it is inherently less constrained and less correlated to the local pose compared to the body pose. Specifically, we initialize the latent code $\mathbf{z}^h$ from the pre-trained encoder. Our objective is to perform the optimization over the latent code $\mathbf{z}^h$, global motion state ${}^{\text{w}}\mathbf{Q}^h$ and the scale factor ω during the optimization by minimizing the following objectives.

$$E_{II}({}^{\text{w}}\mathbf{Q}^h, \omega, \mathbf{R}_t, \boldsymbol{T}_t) = \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{pen}} + \mathcal{L}_{\text{bio}} + \lambda_{2\text{d}}\mathcal{L}_{2\text{d}} + \lambda_s\mathcal{L}_{\text{smooth}} + \lambda_{\text{cam}}\mathcal{L}_{\text{cam}} + \lambda_{\boldsymbol{J}}\mathcal{L}_{\boldsymbol{J}} + \lambda_{\boldsymbol{\beta}}\mathcal{L}_{\boldsymbol{\beta}}. \quad (7)$$

Prior loss ($\mathcal{L}_{\text{prior}}$) We define $\mathcal{L}_{\text{prior}} = \lambda_{\mathbf{z}}\mathcal{L}_{\mathbf{z}} + \lambda_{\boldsymbol{\phi}}\mathcal{L}_{\boldsymbol{\phi}} + \lambda_{\tau}\mathcal{L}_{\tau}$, where $\mathcal{L}_{\mathbf{z}}$ ensures that the motion falls in the motion prior by penalizing the negative log-likelihood:

$$\mathcal{L}_{\mathbf{z}} = \sum_{h\in\{l,r\}} \sum_{t=0}^{T} -\log\mathcal{N}(z^h; \mu^h(\{\boldsymbol{J}_t^h\}), \sigma^h(\{\boldsymbol{J}_t^h\})). \quad (8)$$

The two other terms ensure an as jitter-free as possible trajectory by encouraging global consistency over the global orientation Φ^h and translation ${}^{\text{w}}\tau$:

$$\mathcal{L}_{\boldsymbol{\phi}} = \sum_{t=0}^{T} d_{\boldsymbol{\phi}}({}^{\text{w}}\boldsymbol{\phi}_t, {}^{\text{w}}\hat{\boldsymbol{\phi}}_t) \quad \text{and} \quad \mathcal{L}_{\tau} = \sum_{t=0}^{T} \|{}^{\text{w}}\tau_t^h - {}^{\text{w}}\hat{\tau}_t^h\|^2. \quad (9)$$

Biomechanical loss ($\mathcal{L}_{\text{bio}}$) While the hand prior helps correcting certain implausible configurations, it is still necessary to explictly the hand pose for improved motion quality. Hence, we further add biomechanical constraints [17] to our objective function, which consists of three terms: $\mathcal{L}_{\text{bio}} = \lambda_{palm}\mathcal{L}_{\text{palm}} + \lambda_{ja}\mathcal{L}_{\text{ja}} + \lambda_{bl}\mathcal{L}_{\text{bl}}$. For i^{th} finger bone, each of the terms is defined as:

$$\mathcal{L}_{\text{ja}} = \sum_i d_{H,\alpha}(\boldsymbol{\alpha}_{1:T}^i, \mathbf{H}^i), \quad (10)$$

$$\mathcal{L}_{\text{bl}} = \sum_i \mathcal{I}(\|\mathbf{b}_{1:T}^i\|_2; b_{\text{min}}^i, b_{\text{max}}^i), \quad (11)$$

$$\mathcal{L}_{\text{palm}} = \sum_i \mathcal{I}(\|\mathbf{c}_{1:T}^i\|_2; c_{\text{min}}^i, c_{\text{max}}^i) + \sum_i \mathcal{I}(\|\mathbf{d}_{1:T}^i\|_2; d_{\text{min}}^i, d_{\text{max}}^i), \quad (12)$$

where $\mathcal{L}_{\text{bl}}$ is for bone length, $\mathcal{L}_{\text{palm}}$ is for palmar region optimization, and $\mathcal{L}_{\text{ja}}$ is for joint angle priors. $\mathcal{L}_{\text{ja}}$ constrains the sequence of joint angles for the i-th finger bone $\boldsymbol{\alpha}_{1:T}^i = (\boldsymbol{\alpha}_{1:T}^f, \boldsymbol{\alpha}_{1:T}^a)$ by approximating the convex hull on $(\boldsymbol{\alpha}_{1:T}^f, \boldsymbol{\alpha}_{1:T}^a)$ plane with the point set $\mathbf{H}^i$, and the objective is to minimize the distance $d_{\boldsymbol{\alpha},\mathbf{H}}$ between them. $\mathcal{I}$ is the interval loss penalizing the outliers, and b_i is the bone length of i-th bone. Finally,$\mathcal{L}_{\text{palm}}$ penalizes the outliers of curvature range $(c_{\text{min}}^i, c_{\text{max}}^i)$ and angular distance range $(d_{\text{min}}^i, d_{\text{max}}^i)$ to constraint for the 4 root bones of palm.

Penetration loss ($\mathcal{L}_{\text{pen}}$) The final loss enhances the reconstruction quality of challenging hand interactions by incorporating a contact and penetration penalty term:

$$\mathcal{L}_{\text{pen}} = \sum_{t=0}^{T} \left(\sum_{\mathbf{v}_t^r \in \mathbf{V}_t^r} \min_{\mathbf{v}_t^l \in \mathbf{V}_t^l} \|\mathbf{v}_t^l - \mathbf{v}_t^r\|^2 + \sum_{\mathbf{v}_t^l \in \mathbf{V}_t^l} \min_{\mathbf{v}_t^r \in \mathbf{V}_t^r} \|\mathbf{v}_t^l - \mathbf{v}_t^r\|^2 \right) \quad (13)$$

where $\mathbf{V}_t^l$ and $\mathbf{V}_t^r$ are the intersected vertices of the predicted left hand and right hand, respectively.

3. Experiments

Implementation details We use the L-BFGS algorithm for our three-stage optimization with learning rate $lr = 1$. For Stage 2, we use $\lambda_{2d} =, \lambda_{2d} = 0.001, \lambda_{smooth} = 10, \lambda_{cam} =, \lambda_{\theta} = 0.04, \lambda_{\beta} = 0.05$. For stage 3, we use $\lambda_z = 200, \lambda_{\phi} = 2, \lambda_{\gamma} = 10, \lambda_{pen} = 10, \lambda_{\beta} = 0.05, \lambda_{ja} = 1, \lambda_{palm} = 1, \lambda_{bl} = 1$.

Evaluation metrics Following [16, 20, 22], we split the sequences into 100 frames per segment and align each sequence with ground truth using the first two frames or the

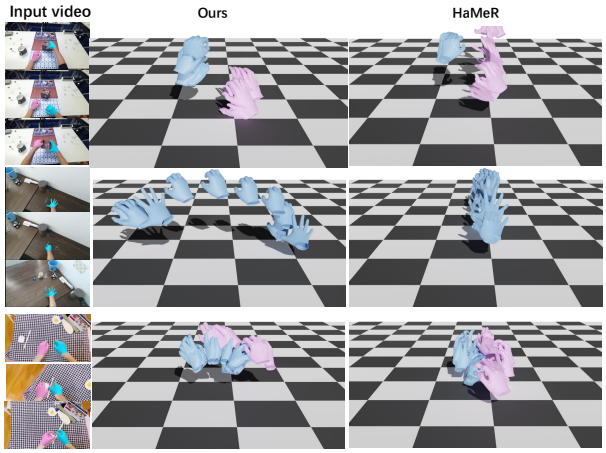

Figure 3. Qualitative comparison with state-of-the-art method HaMeR [14]. (a) is from H2O dataset [7], and (b)(c) are from HOI4D dataset [10] and in-the-wild videos.

Method	G-MPJPE (↓)	GA-MPJPE (↓)	MPJPE (↓)	Acc Err (↓)
ACR [21]	105.6	86.7	47.5	14.2
IntagHand [9]	99.3	73.5	45.8	13.3
HaMeR [14]	91.5	69.5	30.6	9.13
Ours (w/o Stage III)	48.6	39.9	25.4	9.7
Ours	**43.5**	**31.2**	**21.5**	**4.1**

Table 1. Quantitative evaluation results for H2O dataset.

whole sequence. For the first group of the evaluation protocols, **local pose and shape evaluation** in camera coordinate system, we report the commonly used Mean Per Joint Position Error (**MPJPE**). We further report Acceleration error (**Accel**, m/s^2) after root alignment to measure the smoothness between frames of the reconstructed motions, which is computed as the difference between the magnitude of the acceleration vector at each joint. In terms of **global motion evaluation**, the errors can accumulate over time with dynamic cameras. We therefore follow prior arts [16, 20, 22] to use a sliding window and split sequences into smaller segments of 100 frames and align each output segment with the ground-truth data using the first two frames (**G-MPJPE**) or the entire segment (**GA-MPJPE**) in the world coordinate system.

Datasets We employ egocentric hand motion datasets in our experiments following the official split to conduct our experiments: 1) **H2O** [7], 2) **HOI4D** [10], 3) **FPHA** [4].

3.1. Global Motion Evaluation

Quantitative results We report results following the official test split for the egocentric dynamic camera views as shown in Tab. 1 and supplementary materials. It can be observed that our method significantly outperforms the state-of-the-art methods [9, 14, 21] in terms of **G-MPJPE** and **GA-MPJPE**, which demonstrates the superiority of our global 4D hand motion recovery. The results of [9, 14, 21] are obtained from running the official checkpoints.

Qualitative results In Fig. 3, we present detailed comparisons with previous state-of-the-art methods, where we can observe that our method has significant improvements on real plausible motion trajectory with a more plausible **depth reasoning** between two hands. Our method is the only one that can recover the authentic global trajectories and keep them consistent the input videos with dynamic cameras, whilst other methods suffer from the ambiguity of depth and absence of camera pose. We also provide more qualitative results in the video in supplemental materials and on the project page. As there is not yet an interacting hand dataset with dynamic cameras, we provide in-the-wild results.

3.2. Hand Reconstruction Evaluation

To evaluate the per-frame hand pose estimation, we conduct a comprehensive comparison of our method and the existing state-of-the-art hand reconstruction methods: HaMeR [14], IntagHand [9] and ACR [21]. In Tab. 1, we compare the reconstruction accuracy of our baseline (w/o stage III) with the state-of-the-art methods on H2O [7] and FPHA [4] dataset with **MPJPE**. Furthermore, we evaluate the inter-frame smoothness using the acceleration error compared to the existing works, where we can observe significant improvements with lower error in our pipeline. Due to the page limit. Please find more details in supplementary materials and further comparison on FPHA [4] dataset.

3.3. Ablation Study

Method	G-MPJPE (↓)	GA-MPJPE (↓)	MPJPE (↓)	Acc Err (↓)
Ours (w/o Stage III)	48.6	39.9	25.4	9.7
Ours (w/o bio. constraints)	47.3	40.2	24.8	4.3
Ours (w/o pen. constraints)	44.1	32.5	21.9	**4.0**
Ours	**43.5**	**31.2**	**21.5**	4.1

Table 2. Ablation study on H2O dataset.

To fully assess the effectiveness of our proposed method, we perform further ablation studies on the pipeline design to analyze the contribution of each component. In particular, we investigate the effectiveness of the key components: (i) the $\mathcal{L}_{bio}$ and (ii) $\mathcal{L}_{pen}$ and (iii) the interacting hand motion prior module in Stage III. It can be seen from Tab. 2 that incorporating Stage III can boost the performance by a considerable margin as it provides a well-learned motion prior information for the final stage optimization and yields more plausible and smoother 4D trajectory reconstructions.

4. Conclusion

We introduced **Dyn-HaMR**, which achieves the state-of-the-art 4D global motion for interacting hands from complex scenarios even with dynamic cameras, leveraging the SLAM systems in conjunction with the hand priors.

References

[1] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019. 1

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 1

[3] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J. Black. Hmp: Hand motion priors for pose and shape estimation from video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6353–6363, 2024. 1, 2

[4] 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 Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[5] Stuart Geman. Statistical methods for tomographic image restoration. *Bull. Internat. Statist. Inst.*, 52:5–21, 1987. 2

[6] Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, and Joey Tianyi Zhou. A2j-transformer: Anchor-to-joint transformer network for 3d interacting hand pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8846–8855, 2023. 1

[7] 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 (ICCV)*, pages 10138–10148, 2021. 4

[8] Jihyun Lee, Minhyuk Sung, Honggyu Choi, and Tae-Kyun Kim. Im2hands: Learning attentive implicit representation of interacting two-hand shapes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[9] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 4

[10] 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 (CVPR)*, pages 21013–21022, 2022. 4

[11] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 2

[12] Gyeongsik Moon. Bringing inputs to shared domains for 3D interacting hands recovery in the wild. In *CVPR*, 2023. 1

[13] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[14] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *arxiv*, 2023. 1, 2, 4

[15] Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017. 1, 3

[16] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. *arXiv preprint arXiv:2312.07531*, 2023. 1, 3, 4

[17] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European conference on computer vision*, pages 211–228. Springer, 2020. 3

[18] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[19] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 1, 2

[20] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3, 4

[21] Zhengdi Yu, Shaoli Huang, Fang Chen, Toby P. Breckon, and Jue Wang. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 4

[22] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11038–11049, 2022. 3, 4

[23] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11354–11363, 2021. 1

[24] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2354–2364, 2019.

[25] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020. 1