Predicting 4D Hand Trajectory from Monocular Videos

Supplementary Material

	Method	New Days			VISOR		
		@0.05	@0.1	@0.15	@0.05	@0.1	@0.15
All Joints	FrankMocap[14]	16.1	41.4	60.2	16.8	45.6	66.2
	METRO[9]	14.7	38.8	57.3	16.8	45.4	65.7
	Mesh Graphormer [8]	16.8	42.0	59.7	19.1	48.5	67.4
	HandOccNet (param) [12]	9.1	28.4	47.8	8.1	27.7	49.3
	HandOccNet (no param) [12]	13.7	39.1	59.3	12.4	38.7	61.8
	HaMeR [13]	48.0	78.0	88.8	43.0	76.9	89.3
	HaPTIC (Ours)	48.6	79.0	89.9	46.8	79.5	90.9
Visible Joints	FrankMocap	20.1	49.2	67.6	20.4	52.3	71.6
	METRO	19.2	47.6	66.0	19.7	51.9	72.0
	Mesh Graphormer	22.3	51.6	68.8	23.6	56.4	74.7
	HandOccNet (param)	10.2	31.4	51.2	8.5	27.9	49.8
	HandOccNet (no param)	15.7	43.4	64.0	13.1	39.9	63.2
	HaMeR	60.8	87.9	94.4	56.6	88.0	94.7
	HaPTIC (Ours)	60.6	88.5	95.0	60.7	89.3	95.5
Occluded Joints	FrankMocap	9.2	28.0	46.9	11.0	33.0	55.0
	METRO	7.0	23.6	42.4	10.2	32.4	53.9
	Mesh Graphormer	7.9	25.7	44.3	10.9	33.3	54.1
	HandOccNet (param)	7.2	23.5	42.4	7.4	26.1	46.7
	HandOccNet (no param)	9.8	31.2	50.8	9.9	33.7	55.4
	HaMeR	27.2	60.8	78.9	25.9	60.8	80.7
	HaPTIC (Ours)	29.4	62.2	80.4	29.7	64.6	83.1

Table 5. Evaluation of single image pose. We report PCK at different thresholds on benchmark HInt [13]. We compare HaPTIC with multiple image-based hand pose estimation baselines.

In supplementary material, we provide additional quantitative comparisons mentioned in the main paper (Sec. A. We also provide further details on implementing network and evaluation metrics (Sec. B). Finally, we visualize more comparisons and results in supplementary videos.

A. Image-based Hand Pose Estimation

As mentioned in Section 4.2, we report quantitative comparisons between HaPTIC general model and other image-based baselines. Table 6 reports 3D pose evaluation and Table 5 reports their 2D projections.

All Splits on HInt HaPTIC outperforms all other baselines in terms of 2D pose alignment on the more appearance-diverse benchmark HInt. The advantage is even more significant on the occluded split. This is likely because our global spatial attention (Global CA) to the original frame provides more context under occlusion.

3D Hand Pose Benchmark HaPTIC performs comparably in terms of 3D hand poses on HO3D (second best across all methods with gaps to the best model less than 1%). It is slightly worse than HaMeR probably because the HO3D dataset is sampled less in the combination of video datasets. Yet, HaPTIC predicts much more realistic 4D trajectories.

Method	**AUC_J ↑**	**PA-MPJPE ↓**	**AUC_V ↑**	**PA-MPVPE ↓**	**F@5 ↑**	**F@15 ↑**
Liu et al. [10]	0.803	9.9	0.810	9.5	0.528	0.956
HandOccNet [12]	0.819	9.1	0.819	8.8	0.564	0.963
I2UV-HandNet [1]	0.804	9.9	0.799	10.1	0.500	0.943
Hampali et al. [4]	0.788	10.7	0.790	10.6	0.506	0.942
Hasson et al. [6]	0.780	11.2	0.777	11.1	0.464	0.939
ArtiBoost [16]	0.773	11.4	0.782	11.4	0.488	0.944
Pose2Mesh [3]	0.754	12.5	0.749	12.7	0.441	0.909
I2L-MeshNet [11]	0.775	11.2	0.722	13.9	0.409	0.932
METRO [9]	0.792	10.4	0.779	11.1	0.484	0.946
MobRecon [2]	-	9.2	-	9.4	0.538	0.957
Keypoint Trans [5]	0.786	10.8	-	-	-	-
AMVUR [7]	0.835	8.3	0.836	8.2	0.608	0.965
HaMeR [13]	0.846	7.7	0.841	7.9	0.635	0.980
HaPTIC (Ours)	0.842	8.0	0.839	8.1	0.628	0.980

Table 6. Evaluation of single image pose. We report quality of 3D hand pose on benchmark HO3D. We compare HaPTIC with multiple image-based hand pose estimation baselines.

B. Implementation Details

Network Implementation Details We use AdamW optimizer with learning rate $1e-4$. We train on eight H100 with batch size 16 (8 for video and 8 for images) for 1000000 iterations. We use the same weights as that in HaMeR to mix data from different image datasets. The (unnormalized) sampled weights to mix video datasets are 0.25 (ARCTIC-EGO), 0.15 (ARCTIC-EXO), 0.05 (DexYCB), 0.05 (H2O), 0.05 (HO3D), 0.15 (InterHand2.6M).

In test time optimization, we use AdamW optimization with learning rate $1e-3$ and optimize for 1000 iterations for each sequence. The optimization objective consists of 2D reprojection loss with ground truth 2D keypoints, acceleration loss of predicted 3D and 2D keypoints, and distance to the original predictions.

Evaluation Metrics. To calculate GA/FA-MPJPE, we align the predicted trajectory with the ground truth by searching for an affine transformation (isometric scale, rotation, and translation) between selected keypoints. In GA-MPJPE, the selected keypoints are all joints from all frames while in FA-MPJPE, they are all joints from the first frame.

In ARCTIC, each video takes about 30 seconds. Following the literature on whole-body reconstruction in global coordinates [15, 17], we clip long sequences into shorter clips of the same length (60) before computing global trajectory metrics.

C. More Results in Videos

We provide more results in Supplementary Videos, including comparisons with other models (Fig. 4), comparisons before and after optimization (Fig. 5), and more qualitative

results from HaPTIC (Fig. 1 and 6). They can be viewed by opening https://judyye.github.io/haptic-www/

References

[1] Ping Chen, Yujin Chen, Dong Yang, Fangyin Wu, Qin Li, Qingpei Xia, and Yong Tan. I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. In *ICCV*, 2021. 1

[2] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *CVPR*, 2022. 1

[3] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, 2020. 1

[4] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3D annotation of hand and object poses. In *CVPR*, 2020. 1

[5] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *CVPR*, 2022. 1

[6] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1

[7] Zheheng Jiang, Hossein Rahmani, Sue Black, and Bryan M Williams. A probabilistic attention model with occlusion-aware texture regression for 3d hand reconstruction from a single rgb image. In *CVPR*, 2023. 1

[8] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021. 1

[9] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 1

[10] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, 2021. 1

[11] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single rgb image. In *ECCV*, 2020. 1

[12] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. HandOccNet: Occlusion-robust 3D hand mesh estimation network. In *CVPR*, 2022. 1

[13] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 1

[14] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3D whole-body pose estimation system via regression and integration. In *ICCVW*, 2021. 1

[15] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *CVPR*, 2024. 1

[16] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *CVPR*, 2022. 1

[17] Paper Authors Your. Slahmr: Simultaneous localization and human mesh recovery. In *CVPR*, 2023. 1