Supplementary Material for
HMP: Hand Motion Priors for Pose and Shape Estimation From Video

1. Additional Experiments

1.1. HO-3D Test Split Results

We report quantitative results over the HO3D test split in Tab. 1. In addition to mean-per-joint-projection error (PA-MPJPE) and vertex-to-vertex error (PA-V2V) we provide the F-scores after procrustes alignment: PA-F@5, and PA-F@15. Those values are obtained from the official evaluation server using the test set of HO-3D. Since we do not have the ground-truth labels for the test set, we cannot compute and report RA-ACC.

Recent methods utilize HO3D and DexYCB as their primary training datasets. However, given the limited background and subject diversity inherent to HO3D and DexYCB, methods solely trained on these datasets struggle to generalize effectively to in-the-wild videos. In contrast, neither PyMAF-X nor our motion prior relies on these datasets for training, thereby enhancing their generalization to in-the-wild scenarios. Consequently, directly comparing our method with those trained on HO3D and DexYCB can be challenging. To signify this distinction, we have marked such methods with † in the corresponding tables. Overall, our method outperforms the existing state-of-the-art (SOTA) techniques on the HO3D-test split. Furthermore, our approach enhances the performance of the PyMAF-X method, which we employ for initialization, across both datasets.

	HO3D-v3			
Methods	**PA-MPJPE ↓**	**PA-V2V 2V ↓**	**PA-F@5 ↑**	**PA-F@15 ↑**
Hasson *et al.* † [4]	11.4	11.4	42.8	93.2
Hasson *et al.* † [5]	11.1	11.0	46.0	93.0
TempCLR† [15]	10.6	10.6	48.1	93.7
Hampali *et al.* † [2]	10.7	10.6	50.6	94.2
Liu *et al.* † [8]	10.1	9.7	53.2	95.2
Deformer† [1]	9.4	9.1	54.6	**96.3**
HandOccNet† [11]	**9.1**	**8.8**	**56.4**	**96.3**
PyMAF-X [13]	11.2	11.0	47.6	92.8
HMP (Ours)	**10.2**	**9.9**	**51.0**	**94.6**

Table 1. State-of-the-art comparison on the test split of HO3D-v3 dataset [3]. Methods denoted with † uses HO-3D as their training dataset.

MP Type	Diversity ↑
PCA-based	5.83
GMM-based	5.79
HMP (Ours)	**5.86**

Table 2. Diversity metrics for different motion prior types

1.2. Sample Diversity of Different Motion Priors

We report the sample diversity metrics for PCA-based motion prior, GMM-based motion prior, and our motion prior in Tab. 2. All motion priors are trained on the same sequences from the AMASS dataset. We follow the same evaluation criteria as [6].

2. Additional Qualitative Results & Remarks

Obtaining 2D Pose Confidence: We use keypoint confidences to weight $\mathcal{L}_{\text{2D}}$. Unfortunately, MediaPipe does not provide separate confidences for hand keypoints [9]. To obtain confidence keypoints, we augment with 11 different views through random rotation and scaling. We perform detection on these views and project the results back. For each joint with index j we compute an std value σ_j:

$$\sigma_j^2 = \frac{1}{N}\sum_{n=1}^{N}(P_n - P_0)^2. \tag{1}$$

Here P_0 and P_n denotes the original view and n'th view obtained by random scaling and rotation. To discard anomalies, we clip the std values by an upper threshold γ:

$$\sigma_j = \min(\sigma_j, \gamma). \tag{2}$$

We then compute confidence value as,

$$\alpha_j = 1 - \frac{\sigma_j}{\gamma}. \tag{3}$$

We set N=11 and γ = 4.

Qualitative Results: We provide more qualitative results on DexYCB dataset and in-the-wild videos in Fig. 1, 2, and 3. We refer the readers to our SupMat video for more results.

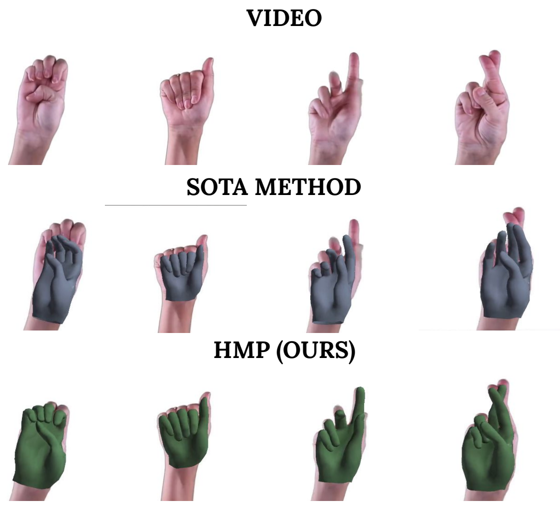

Figure 1. **3D hand pose and shape estimation on an in-the-wild video:** input video (top), PyMAF-X (middle), HMP (bottom)

Hyperparameter	Value
Epochs	1000
Batch Size	16
Learning Rate	1e-4
Weight Decay	1e-4
Latent Dimension	1024

Table 3. Hyperparameter setting in motion prior training

3. Training A Hand Motion Prior

Data Preprocessing: We adopted a similar approach to data preprocessing as [6, 12]. GRAB, TCDHands, and SAMP datasets in AMASS have hand articulation [10]. We only train motion prior for the right hand. We first reflect left hand articulations in the dataset for data augmentation. Then all sequences are divided to motion clips of 128 timesteps. These clips are preprocessed to obtain processed data $\mathbf{X}$. For any timestep t, the processed data point $\mathbf{X}_t$ is,

$$\mathbf{X}_t = \left(\mathbf{x}_t^p \quad \dot{\mathbf{x}}_t^p \quad \mathbf{x}_t^r \quad \dot{\mathbf{x}}_t^r \right) \in \mathbb{R}^{J \times 15}. \tag{4}$$

For a timestep t, $\mathbf{x}_t^r \in \mathbb{R}^{J\times3}$ denotes joint positions, $\dot{\mathbf{x}}_t^p \in \mathbb{R}^{J\times3}$ denotes joint velocities, $\mathbf{x}_t^r \in \mathbb{R}^{J\times6}$ denotes hand pose in 6D rotation representation [14], $\dot{\mathbf{x}}_t^r \in \mathbb{R}^{J\times3}$ denotes angular velocity. J indicates the number of joints.

Architecture: The architecture is adapted from [6]. We employed the Adam optimizer [7]. The training process takes ∼10 hours on NVIDIA-A100 GPU. Our motion prior contains parameters contains ∼63.4 million parameters. Please refer to Table 3 for hyperparameter values.

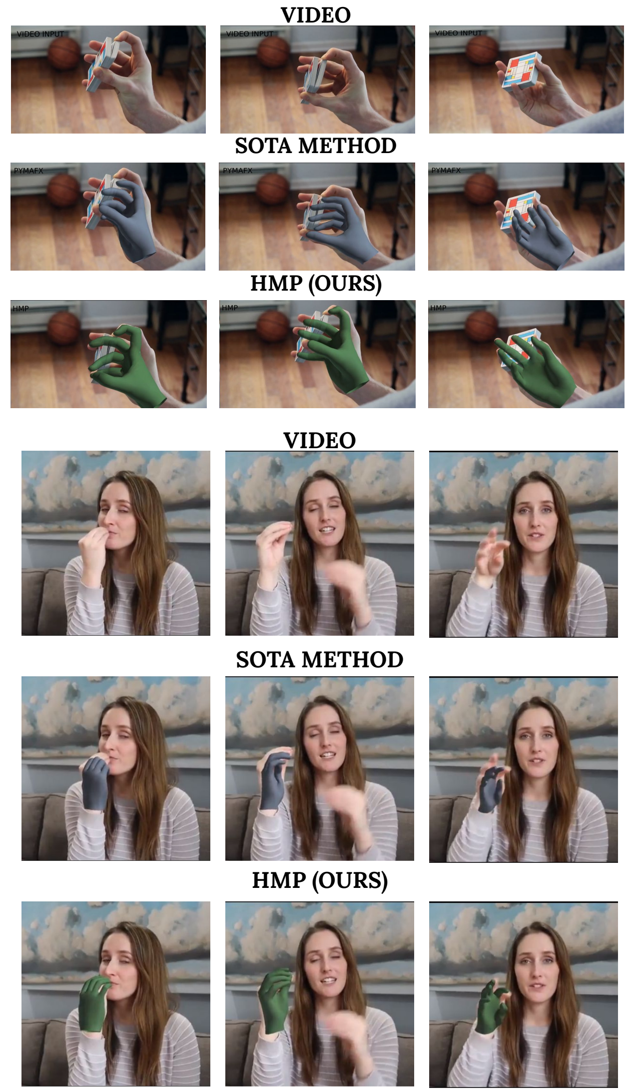

Figure 2. **3D hand pose and shape estimation on an in-the-wild video:** input video (top), PyMAF-X (middle), HMP (bottom)

4. Failure Cases

Keypoint Detection Failure: One key cause for our method's failure is inaccurate keypoints. Under motion blur and occlusion, current state-of-the-art keypoint detectors tend to fail providing correct detections. Fig. 5 shows those cases along with our methods' output.

Bounding Box Discontinuity: Another risk of failure originates from hand bounding box detection. Our method fails to interpolate and perform motion inbetweening. This usually happens with motion blur. An example can be seen in Fig. 4.

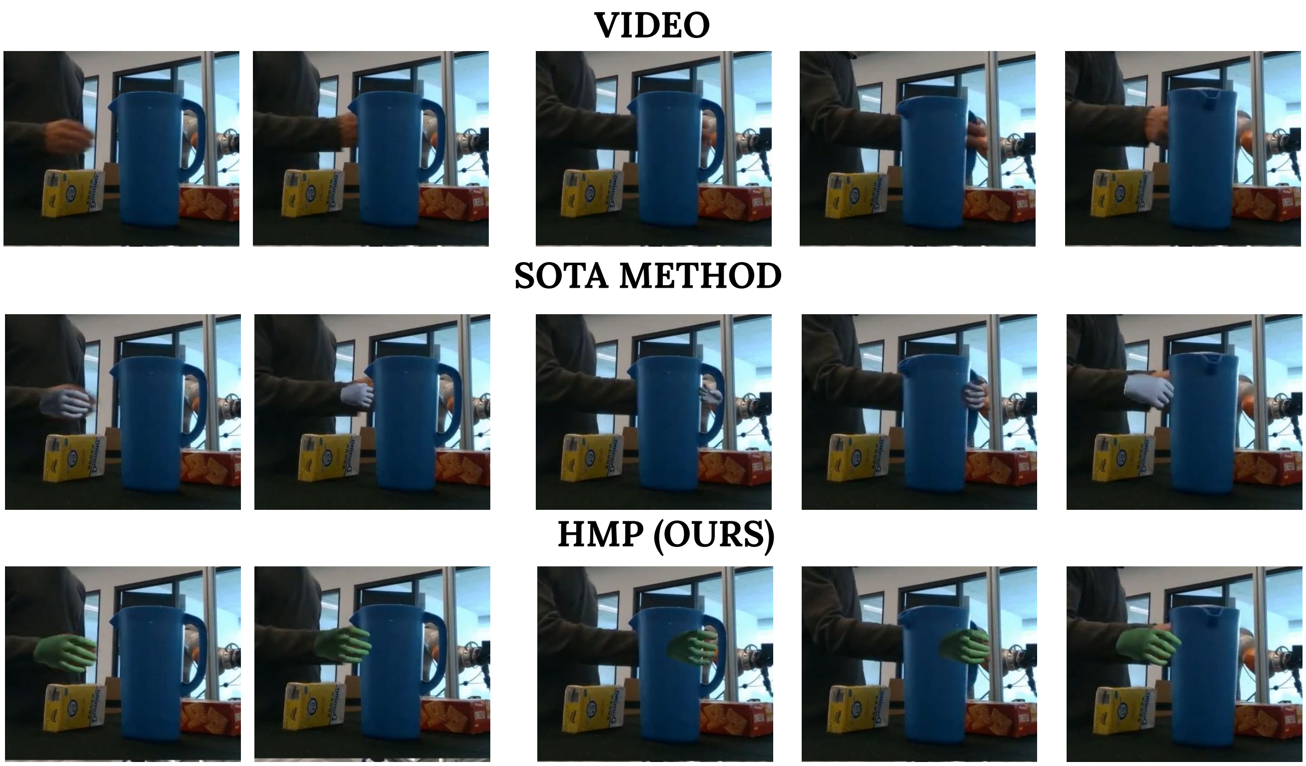

Figure 3. **3D hand pose and shape estimation on DexYCB videos:** input video (top), PyMAF-X (middle), HMP (bottom)

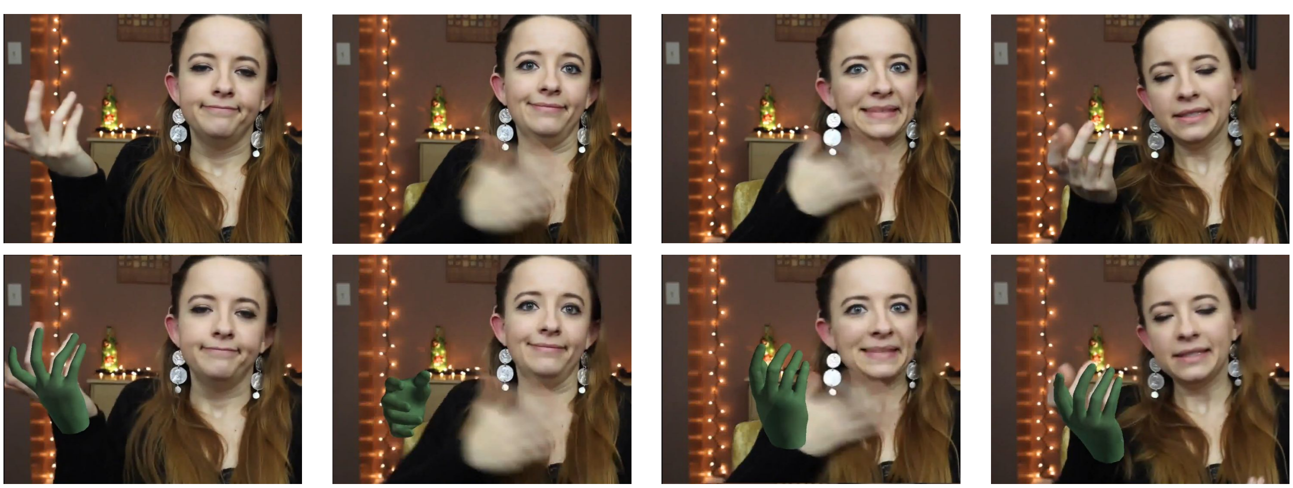

Figure 4. Failure in motion inbetweening due to the discontinuity in bounding box detection. Bounding boxes are detected only for the first and the final frame.

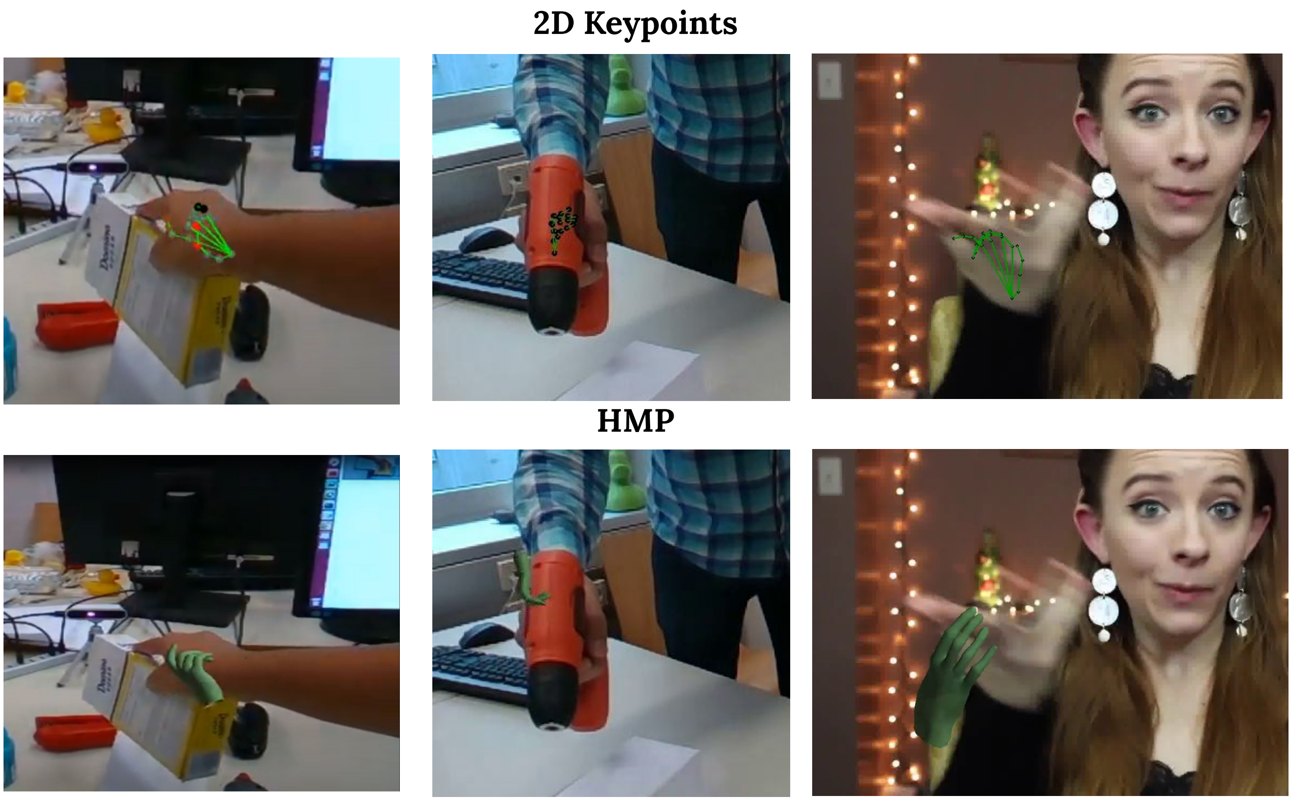

Figure 5. Failure cases caused by faulty keypoint detections. Keypoints detected (top), HMP (bottom)

References

[1] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. *ArXiv*, abs/2303.04991, 2023. 1

[2] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3D annotation of hand and object poses. In *CVPR*, pages 3196–3206, 2020. 1

[3] Shreyas Hampali, Sayan Deb Sarkar, and Vincent Lepetit. HO-3D-v3: Improving the accuracy of hand-object annotations of the HO-3D dataset, 2021. 1

[4] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, pages 571–580, 2020. 1

[5] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1

[6] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. NeMF: Neural motion fields for kinematic animation. In *NeurIPS*, 2022. 1, 2

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 2

[8] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3D hand-object poses estimation with interactions in time. In *CVPR*, pages 14687–14697, 2021. 1

[9] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A framework for building perception pipelines, 2019. 1

[10] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 2

[11] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3D hand mesh estimation network. In *CVPR*, pages 1496–1505, 2022. 1

[12] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3d human motion model for robust pose estimation. In *ICCV*, 2021. 2

[13] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE TPAMI*, 2023. 1

[14] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019. 2

[15] Andrea Ziani, Zicong Fan, Muhammed Kocabas, Sammy Christen, and Otmar Hilliges. TempCLR: Reconstructing hands via time-coherent contrastive learning. In *3DV*, 2022. 1