JAN 4-8 WACV 2024 WAIKOLOA HAWAII

HMP: Hand Motion Priors for Pose and Shape Estimation from Video

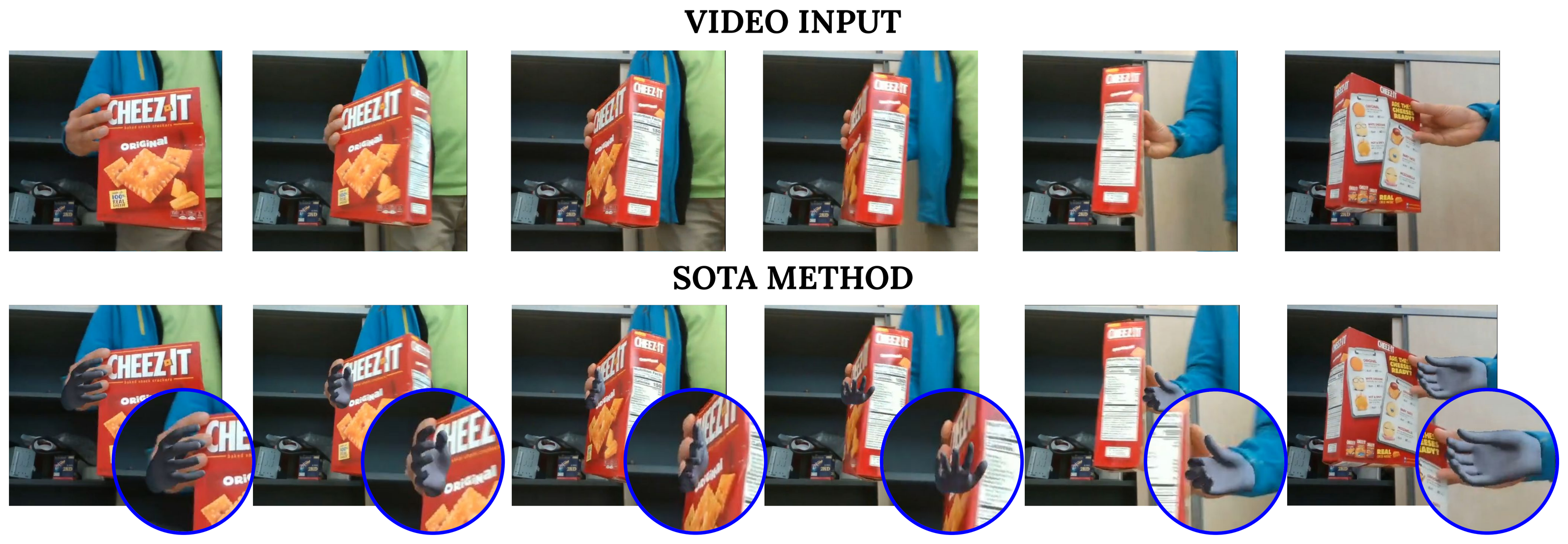
Enes Duran[1,2], Muhammed Kocabas[1,3], Vasileios Choutas[1,3], Zicong Fan [1,3], Michael J. Black[1]

[1]Max Planck Institute for Intelligent Systems [2]University of Tübingen, [3]ETH Zürich

PERCEIVING SYSTEMS MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

imprs-is

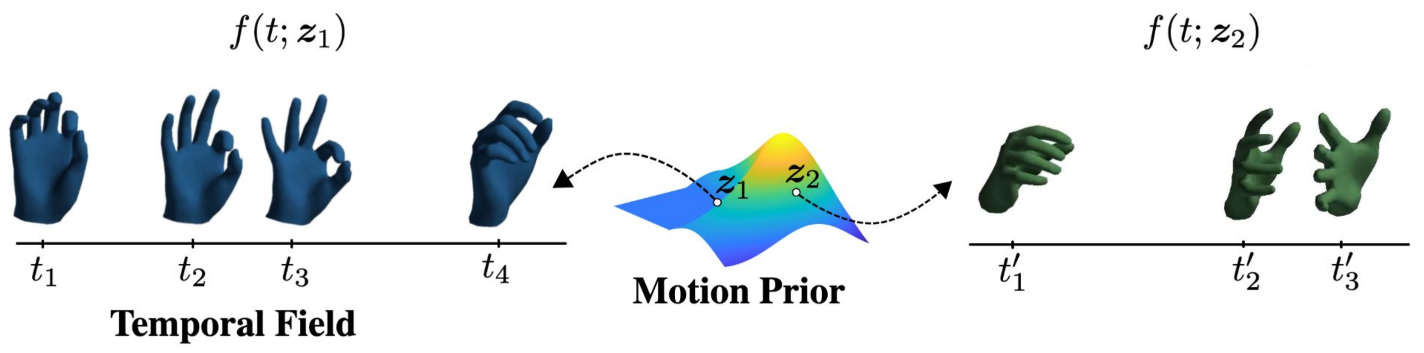
1) Motivation



- **Goal:** To regress hand pose and shape from video.
- **Problem 1:** Current image-based methods do not work well enough on temporal data [1, 2]. They are jittery and susceptible to occlusions.
- **Problem 2:** We do not have enough temporal data to train a direct feed-forward model with high generalization capacity [3, 4, 5].
- **Idea:** Can we use clean (non-video) 3D hand motion-capture data available [6]?

2) Hand Motion Prior

- For hand motion prior we adopt NeMF-based architecture. It is a motion VAE model that represent motion as continuous field [7].



$$\mathcal{E}: \mathbf{X}_t \rightarrow \mathbf{z}_\theta \qquad \mathcal{D}: (t, \mathbf{z}_\theta) \rightarrow \tilde{x}_t \qquad \mathbf{X}_t = (\mathbf{x}^p_t, \dot{\mathbf{x}}^p_t, \mathbf{x}^r_t, \dot{\mathbf{x}}^r_t) \in \mathbb{R}^{J\times 15}$$

- The loss function consists of KL divergence and reconstruction error.

$$\mathcal{L}_{\text{rec}} = \lambda_{\text{rot}}\mathcal{L}_{\text{rot}} + \lambda_{\text{ori}}\mathcal{L}_{\text{ori}} + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}$$

- We use GRAB, TCDHands and SAMP datasets from AMASS [6]. (800K frames)

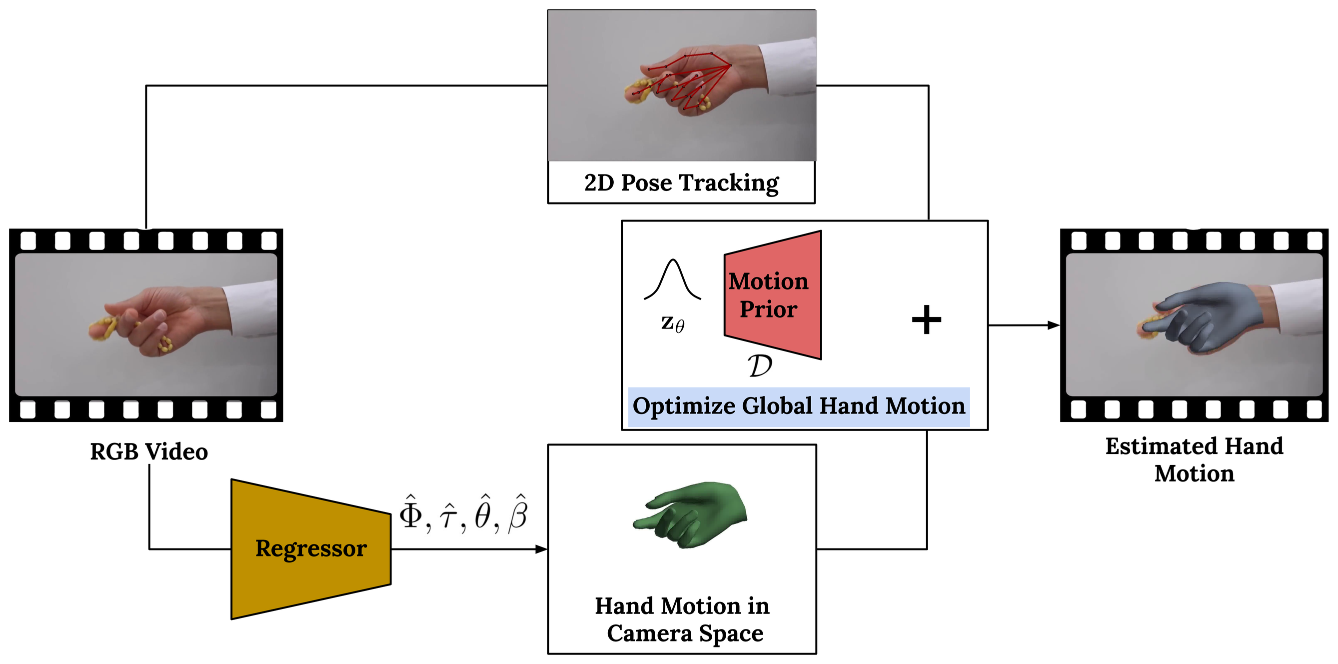
3) Keypoint Blending & Confidence

- MediapPipe has higher accuracy and sparse detection, PyMAF-X 2D projections are dense but not accurate. Therefore, we use MediaPipe if there is a hand detected else projected PyMAF-X 2D keypoints.
- MediaPipe does not provide per-joint confidence. We compute a confidence based on keypoint detection from N augmented views of an image:

- We then project detections back to the original space and calculate standard deviation to approximate the confidence per joint :

$$\sigma_j^2 = \frac{1}{N}\sum_{n=1}^{N}(P_n - P_0)^2, \qquad \sigma_j = \min(\sigma_j, \gamma), \qquad \alpha_j = 1 - \frac{\sigma_j}{\gamma}.$$

4) Method Overview



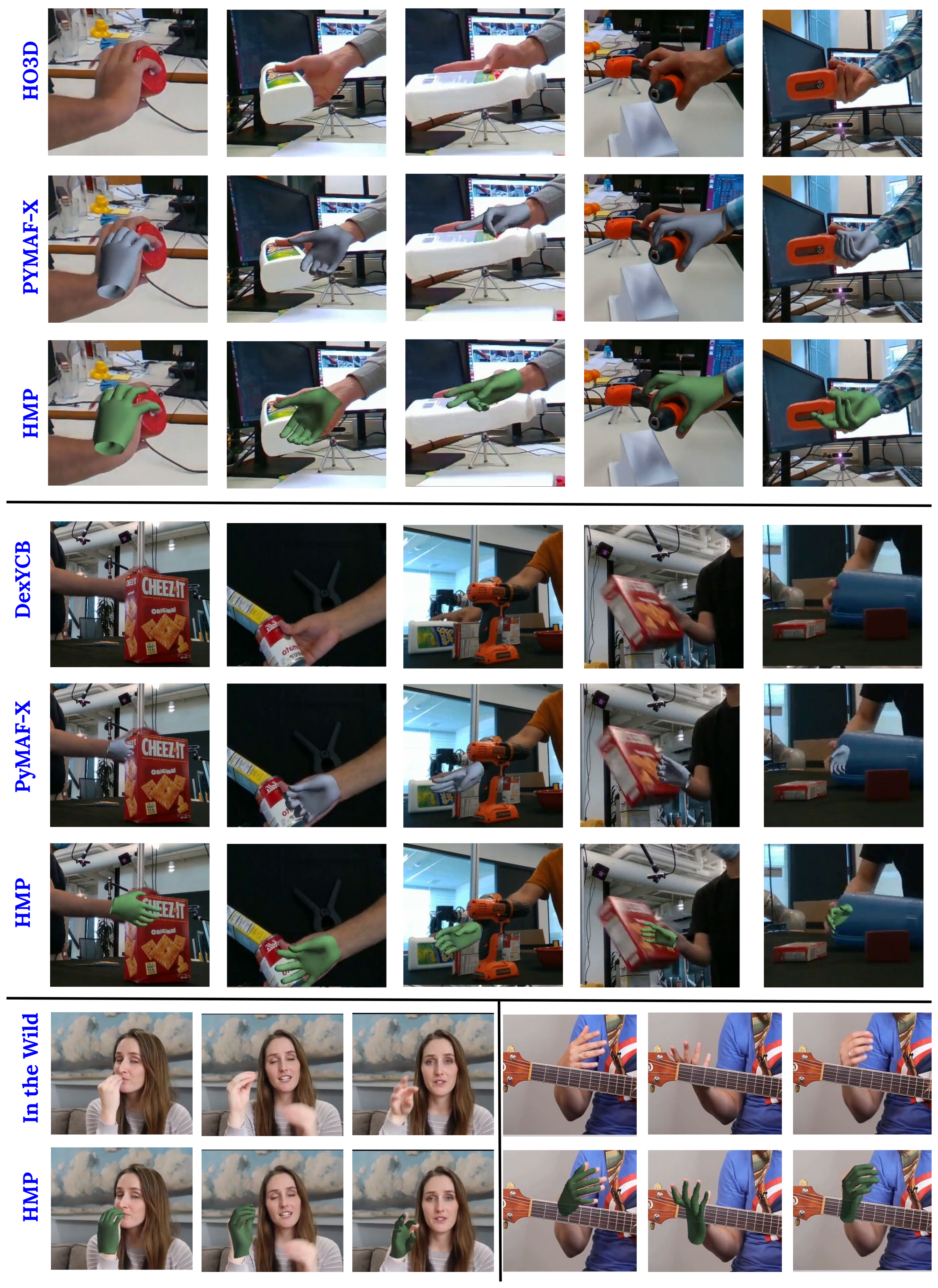
Stg.	Variables	Loss Function	Loss Coefficients
I	Φ, τ, β	$\mathcal{L}_o, \mathcal{L}_{\text{tr}}, \mathcal{L}_\beta, \mathcal{L}_{os}, \mathcal{L}_{ts}, \mathcal{L}_{\text{2D}}$	$lr = 0.05, \lambda_o = 3, \lambda_{\text{tr}} = 1, \lambda_{os} = 1, \lambda_{ts} = 5, \lambda_\beta = 3, \lambda_{\text{2D}} = 0.05$
II	$\Phi, \tau, \beta, \mathbf{z}_\theta$	$\mathcal{L}_o, \mathcal{L}_{\text{tr}}, \mathcal{L}_\beta, \mathcal{L}_{os}, \mathcal{L}_{\text{2D}}, \mathcal{L}_{\text{MP}}$	$lr = 0.05, \lambda_o = 2, \lambda_{\text{tr}} = 1, \lambda_{os} = 1, \lambda_\beta = 10, \lambda_{\text{2D}} = 0.05, \lambda_{\text{MP}} = 300$

- Φ, τ are the global rotation and translation, respectively. Loss terms are:

$$\mathcal{L}_{\text{o}} = \sum_{t=0}^{T} g(\Phi_t, \hat{\Phi}_t)^2, \mathcal{L}_{\text{tr}} = \sum_{t=0}^{T} \|\tau_t - \hat{\tau}_t\|_2^2, \mathcal{L}_{\text{os}} = \sum_{t=0}^{T-1} g(\Phi_{t+1}, \Phi_t)^2, \mathcal{L}_{\text{ts}} = \sum_{t=0}^{T-1} \|\tau_{t+1} - \tau_t\|_2^2,$$

$$\mathcal{L}_{\text{2D}} = \sum_{i=1}^{21} \sum_{t\in T_{\text{detect}}} \alpha_t^i \rho \left(\Pi\left(R_{\text{cam}} J_t^i + T_{\text{cam}}\right) - \mathrm{x}_t^i\right), \mathcal{L}_{\text{MP}} = -\log \mathcal{N}\left(\mathbf{z}_\theta; \mu_\theta\left(\{\theta_t\}\right), \sigma_\theta\left(\{\theta_t\}\right)\right).$$

5) Qualitative Results



6) Quantitative Results & Ablation Studies

- We ablate usage of motion prior and keypoint source selection:

Methods	HO3D-v3 PA-MPJPE ↓	RA-MPJPE ↓	RA-ACC ↓
PyMAF-X [1] + SLERP	10.7	29.4	5.9
No Motion Prior	10.5	28.0	1.9
GMM-based Prior	10.4	27.5	3.4
Stage-1 (PyMAF-X)	10.5	26.8	2.0
Stage-1 (MediaPipe)	10.3	27.0	1.9
Stage-1 (MMPose)	10.3	27.1	**1.8**
Stage-1 (Blend)	10.2	27.7	1.9
Stage-2 (Blend)	**10.1**	**26.7**	2.2
PyMAF-X [1]	10.8	29.6	9.3
PyMAF-X [1] + HMP	**10.1**	**26.7**	2.2
METRO [2]	12.1	38.7	17.4
METRO [2] + HMP	10.8	31.3	2.4

- We report quantitative results on DexYCB [3].

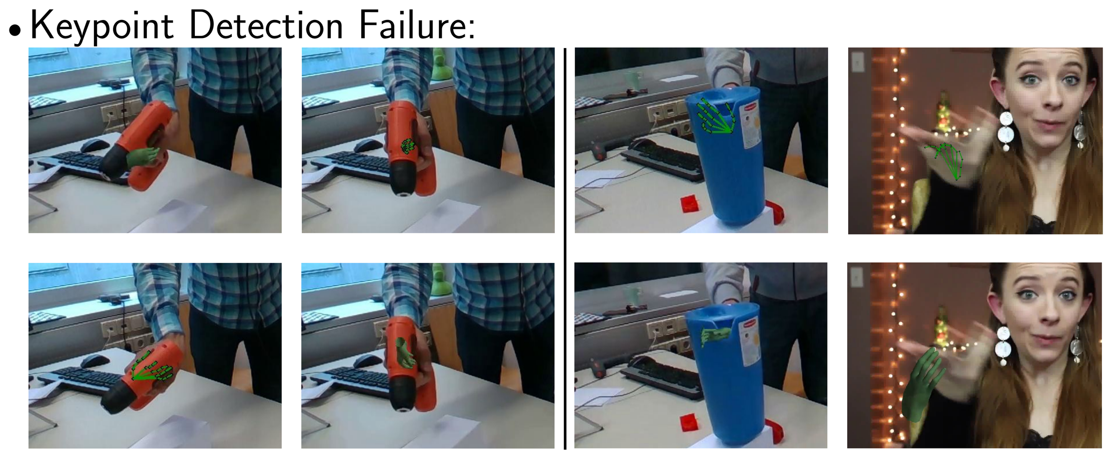
Methods	DexYCB PA-MPJPE ↓	RA-MPJPE ↓	RA-ACC ↓
ArtiBoost† [8]	-	12.8	-
Deformer† [9]	5.2	-	-
PyMAF-X [1]	11.6	38.1	17.1
[1] + HMP(Ours)	**8.9**	**34.1**	**3.6**

- We test HMP on occlusion-specific train subset of HO3D [4]

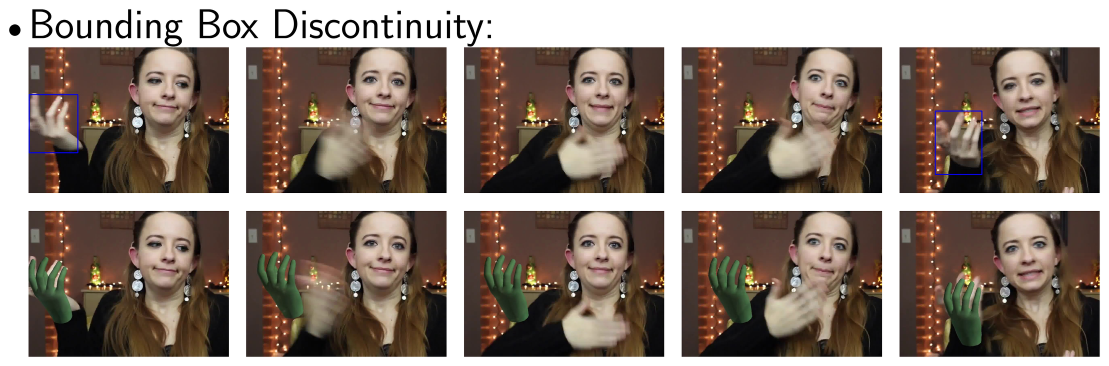
Methods	HO3D-OCC PA-MPJPE ↓	RA-MPJPE ↓	RA-ACC ↓
PyMAF-X [1]	15.3	48.9	26.0
PyMAF-X [1] + SLERP	14.4	41.3	7.9
Stage-1	13.0	38.2	**2.8**
Stage-2	**12.6**	**38.1**	3.0

7) Failure Cases

- Keypoint Detection Failure:



- Bounding Box Discontinuity:



References

[1] Hongwen Zhang et al. "PyMAF-X: Towards Well-aligned Full-body Model Regression from Monocular Images". In: *IEEE TPAMI* (2023).
[2] Kevin Lin, Lijuan Wang, and Zicheng Liu. "End-to-End Human Pose and Mesh Reconstruction with Transformers". In: *CVPR*. 2021.
[3] Yu-Wei Chao et al. "DexYCB: A benchmark for capturing hand grasping of objects". In: *CVPR*. 2021, pp. 9044–9053.
[4] Shreyas Hampali et al. "Honnotate: A method for 3D annotation of hand and object poses". In: *CVPR*. 2020, pp. 3196–3206.
[5] Zicong Fan et al. "ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation". In: *CVPR*. 2023.
[6] Naureen Mahmood et al. "AMASS: Archive of Motion Capture as Surface Shapes". In: *Proceedings of the IEEE International inproceedings on Computer Vision (ICCV)*. 2019.
[7] Chengan He et al. "NeMF: Neural Motion Fields for Kinematic Animation". In: *NeurIPS*. 2022.
[8] Lixin Yang et al. "ArtiBoost: Boosting Articulated 3D Hand-Object Pose Estimation via Online Exploration and Synthesis". In: *CVPR*. 2022.
[9] Qichen Fu et al. "Deformer: Dynamic Fusion Transformer for Robust Hand Pose Estimation". In: *ArXiv* abs/2303.04991 (2023).