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

Enes Duran[1, 2] Muhammed Kocabas[1, 3] Vasileios Choutas[1, †] Zicong Fan[1, 3] Michael J. Black[1]
[1]MPI for Intelligent Systems, Tübingen, Germany [2]University of Tübingen, Germany [3]ETH Zurich, Switzerland
[*]Corresponding author: enes.duran@tue.mpg.de

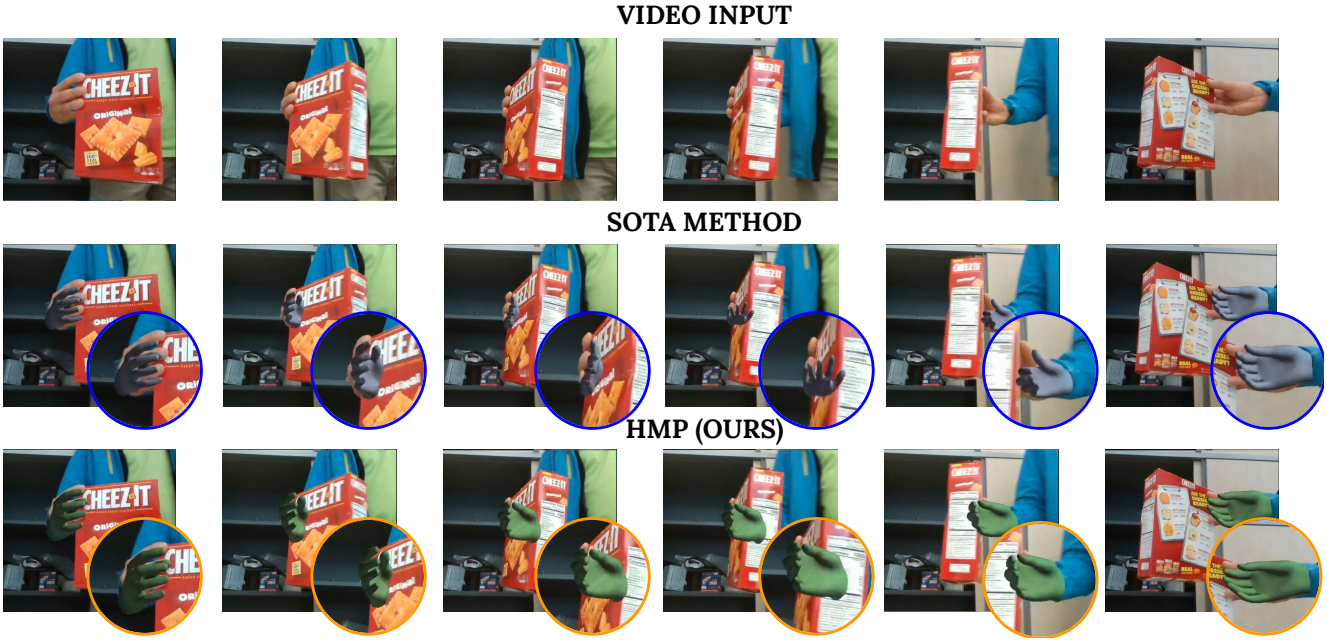

Figure 1. Given challenging hand interaction videos (top), a recent state-of-the-art hand pose estimation approach [64] (middle), fails to produce accurate 3D hand poses. To address this, we exploit a large-scale motion-capture dataset AMASS [37] to train a motion prior and use latent optimization to recover hand pose from videos. Our model HMP (bottom) is robust to occlusion and produce temporally stable results, outperforming previous work on standard benchmarks.

Abstract

Understanding how humans interact with the world necessitates accurate 3D hand pose estimation, a task complicated by the hand's high degree of articulation, frequent occlusions, self-occlusions, and rapid motions. While most existing methods rely on single-image inputs, videos have useful cues to address aforementioned issues. However, existing video-based 3D hand datasets are insufficient for training feedforward models to generalize to in-the-wild scenarios. On the other hand, we have access to large human motion capture datasets which also include hand motions, e.g. *AMASS. Therefore, we develop a generative motion prior specific for hands, trained on the AMASS dataset which features diverse and high-quality hand motions. This motion prior is then employed for video-based 3D hand motion estimation following a latent optimization approach. Our integration of a robust motion prior significantly enhances performance, especially in occluded scenarios. It produces stable, temporally consistent results that surpass conventional single-frame methods. We demonstrate our method's efficacy via qualitative and quantitative evaluations on the HO3D and DexYCB datasets, with special emphasis on an occlusion-focused subset of HO3D. Code is available at* `https://hmp.is.tue.mpg.de`

1. Introduction

Hands often serve as our primary mean for manipulating objects and engaging with our surrounding environments. Therefore, accurately reconstructing the 3D poses and shapes of hands from RGB images plays a crucial

[†]Work done at MPI, now at Google

role in a range of applications including human–computer interaction, augmented/virtual reality (AR/VR), robotics, biomechanics, and animation. Despite years of research in this direction, this task is still challenging due to the high degree of articulation, occlusion caused by hand–object interactions, self-occlusion, and rapid motion inherent to hand movements.

Existing methodologies predominantly investigate the estimation of hand pose and shape from single images [5, 6, 11, 26, 33, 34, 40, 42, 48, 59, 70]. However, such approaches tend to generate temporally-inconsistent reconstructions of hand motion. They are often plagued by jitter, missing predictions, and produce noisy motion results (see middle row of Fig. 1).

In contrast to the aforementioned scenarios, videos serve as a rich source of data for hand motion analysis. Unlike single images, videos contain temporal information that can help to predict coherent hand reconstructions throughout time by learning correlations between time-adjacent frames. They encapsulate a wealth of cues related to hand motion that could improve hand pose and shape estimation. However, this valuable aspect remains largely under-explored, with very few attempts [14, 21, 57, 69] being made to leverage video data. Most recently, Fu *et al*. [14] introduced a feedforward model which takes a video as input and reconstructs the observed hand sequence. However, this method has very limited generalization capability since existing video-based 3D hand motion datasets [8, 17] are limited in terms of the number of subjects and background diversity. On the other hand, large motion capture datasets *e.g*. AMASS [37] contain accurate and diverse 3D hand pose annotation, but they do not contain images.

Motivated by these observations, our key insight is that we can leverage existing MoCap datasets to build a robust, generative 3D hand motion prior and use this generative motion prior for 3D hand pose and shape estimation from monocular videos. After training on the large AMASS motion capture dataset, we use HMP (Hand Motion Priors) as a motion prior at test time for 3D hand pose and shape estimation from noisy and partial observations *e.g*. RGB videos and 2D or 3D joint sequences. In particular, we introduce a latent optimization framework which interacts with HMP to estimate the parameters of hand motion. This interaction happens by parameterizing the motion in the latent space of HMP, and by using HMP priors in order to regularize the optimization towards the space of plausible motions.

Experimental results for 3D hand pose and shape estimation on two existing hand pose estimation video datasets, DexYCB [8] and HO3D [17], show that our method outperforms state-of-the-art methods. To analyze our method's robustness to occlusion, we curate an occlusion-heavy test set from HO3D which we name HO3D-OCC and demonstrate that our approach is more robust to occlusions than existing approaches. Further, we show that our method surpasses traditional motion priors *e.g*. Gaussian Mixture Models, PCA-based temporal priors, as well as a direct optimization hand pose and shape parameters.

In summary, our contributions include:

- We introduce a generative hand motion prior learned from a large-scale MoCap dataset AMASS [37].
- We present a latent optimization-based method for accurate hand pose and shape estimation from monocular videos.
- We demonstrate that our method reconstructs more accurate 3D hand motion under partial or heavy occlusions thanks to our robust generative motion prior. We highlight this on HO3D-OCC, an occlusion-specific subset of HO3D dataset.
- We show that our framework allows us to perform better hand reconstruction results compared to traditional temporal priors or direct optimization of hand pose and shape.

2. Related Work

2.1. Hand Pose Estimation From a Single Image

Methods estimating 3D hand pose from single images can be split into *model-free* and *model-based* approaches.

Model-free methods [33, 34, 42, 48, 59, 70] directly estimate the hand pose by predicting 3D joint positions [6, 10, 50, 52, 70] or joint heatmaps [5, 11, 26, 40]. For instance, Zimmermann *et al*. [70] propose the first convolutional network to detect 2D hand joints and lift them into the 3D space with an articulation prior. Iqbal *et al*. [26] introduce a 2.5D representation allowing to make use of supplementary depth supervision. Likewise, Spurr *et al*. [48] present biomechanical constraints to refine the pose predictions on 2D supervised data. These methods require abundance of annotated data to train due to the lack of 3D priors.

MANO [46] is a parametric model of hands. In MANO, the hand is parameterized by pose and shape parameters. The pose parameters define the articulation of the hand, including finger bending and other movements, while the shape parameters define the overall structure and morphology of the hand. Several model-based approaches [2, 4, 11, 12, 22, 35, 67] directly predict MANO parameters. The 3D hand joint and mesh vertex coordinates are computed from the MANO parameters using linear blend skinning. Zhang *et al*. [67] introduce a framework that harnesses a differentiable re-projection loss for accurate hand mesh recovery. Similarly, Hasson *et al*. [22] use a contact loss that ensures the interaction between the hand and any object appears realistic in predictions. Despite the notable advances achieved by image-based techniques, their results

are still not temporally consistent due to occlusions and motion blur present in single frames.

2.2. Temporal Hand Pose Estimation

Recent methods [21, 35, 41, 49, 57, 69] attempt to leverage temporal data from videos to improve the hand pose estimation performance. Hasson *et al*. [21] use the photometric consistency between the re-projected 3D hand predictions and the optical flow of adjacent frames as supervision. Liu *et al*. [35] train an initial model on an annotated dataset, and deploy it on a large-scale video dataset to collect pseudo-labels. They use pseudo-labels to train a single frame model. Meanwhile, Ziani *et al*. [69] use temporal constrastive learning to learn features that are robust to occlusion and motion blur. They also demonstrate the performance of learned features on a temporal model similar to VIBE [32], a full-body human pose and shape estimation method. Fu *et al*. employ a transformer-driven architecture to process temporal relationship between the input video frames [14], resulting in a temporally coherent and accurate hand pose estimation. A limitation of this method is their reliance on video hand pose datasets for training, but the limited subject and background diversity in current datasets impedes the generalizability of methods which take video as input.

In contrast, our method makes use of AMASS motion capture dataset to learn a robust motion prior and fits the latent code of this motion prior to 2D hand keypoint estimations estimated by off-the-shelf algorithms. This makes it more general and flexible compared to existing works relying on video inputs.

2.3. Motion Prior Models

Given the absence of motion prior models specific for hands, we turn our attention to methods focused on modeling human body movements. There has been a significant amount of research on 3D human dynamics for various tasks, including motion prediction and synthesis [1, 3, 7, 13, 16, 20, 27, 38, 43, 44, 54, 56, 60–62, 68]. Recently, human pose estimation methods have started to incorporate learned human motion priors to help resolve pose ambiguity [32, 45, 66]. Motion-infilling approaches have also been proposed to generate complete motions from partially observed motions [19, 24, 28, 29]. Diffusion models [47] have also been used as priors for motion synthesis and infilling [25,53,63,65]. Rempe *et al*. [45] train an autoregressive VAE-based motion prior on AMASS dataset, called HuMoR. They use HuMoR as a motion prior at test time for 3D human perception from noisy and partial observations across different input modalities such as RGB videos and 2D/3D joints. It is computationally expensive to perform latent optimization since HuMoR is autoregressive. Neural Motion Fields (NeMF) express human motion as a time-conditioned continuous function and demonstrate superior motion synthesis performance [23]. Our approach extends NeMF by leveraging it as a motion prior for hands.

3. Method

Our method HMP consists of two phases (Fig. 2): In the initialization phase, it detects hand bounding boxes, 2D hand keypoints, and initialize MANO hand pose and shape estimates (Sec. 3.2) from video frames. In the multi-stage optimization phase (Sec. 3.4), it then refines those estimates in a video by enforcing hand motion prior constraints.

3.1. Preliminaries

HMP takes as input a video of T frames $I = \{I_1, ..., I_T\}$. The camera is assumed to be static, *i.e*. $\mathbf{R}_{\text{cam}} = \mathbb{I}$ and $\mathbf{T}_{\text{cam}} = [0, 0, 0]$ where $\mathbf{R}_{\text{cam}} \in SO(3), \mathbf{T}_{\text{cam}} \in \mathbb{R}^3$. The hand motion in global coordinate system $\mathbf{Q} = \{Q_t = \{\Phi_t, \tau_t, \theta_t, \beta\}\}_{t=0}^{T}$ consists of global orientation $\Phi_t \in SO(3)$, global translation $\tau_t \in \mathbb{R}^3$, hand pose $\theta_t \in \mathbb{R}^{15\times3}$, and hand shape $\beta \in \mathbb{R}^{10}$ for all visible timesteps t. We use MANO model to represent hand meshes in time [46]. Similar to parametric body models, MANO model outputs a triangulated hand mesh $V_t \in \mathbb{R}^{778\times3}$ for each timestep t derived from hand motion $\mathbf{Q}$.

Existing 3D hand pose datasets such as HO3D [17], DexYCB [8] contain images with hands, but they do not have sufficient data with diverse and accurate 3D hand pose annotation. On other hand, large-scale motion capture datasets such as AMASS [37] have highly-diverse 3D hand motion data captured in accurate mocap setups, but they do not contain images. Our key insight is to leverage large-scale motion capture datasets to address the data scarcity problem. We first train a hand motion model on AMASS, which learns a prior model on natural hand motion. We then introduce a novel optimization-based framework for recovering hand motion by leveraging this motion prior (Sec. 3.4). In this formulation, our method can work with any pose regressors and 2D hand keypoint estimators in plug-and-play fashion (Tabs. 3 and 5).

3.2. Initialization

We first obtain bounding boxes for hands using an off-the-shelf hand tracking model [9]. For each bounding box, we estimate hand pose and shape in global coordinates $\hat{\mathbf{Q}} = \{\hat{Q}_t\}_{t=0}^{T}$, where $\hat{Q}_t = \{\hat{\Phi}_t, \hat{\tau}_t, \hat{\theta}_t, \hat{\beta}_t\}$, using PyMAF-X [64]. Hand bounding box detection methods often fails when the hand is occluded by objects or during two-hand interactions. To obtain initialization for such frames we perform spherical interpolation (SLERP) for the cases where the bounding box confidence is lower than a threshold. Our optimization starts from this initial condition.

For 2D keypoints, we use keypoints combined from MediaPipe and PyMAF-X [36,64]. If MediaPipe does not have

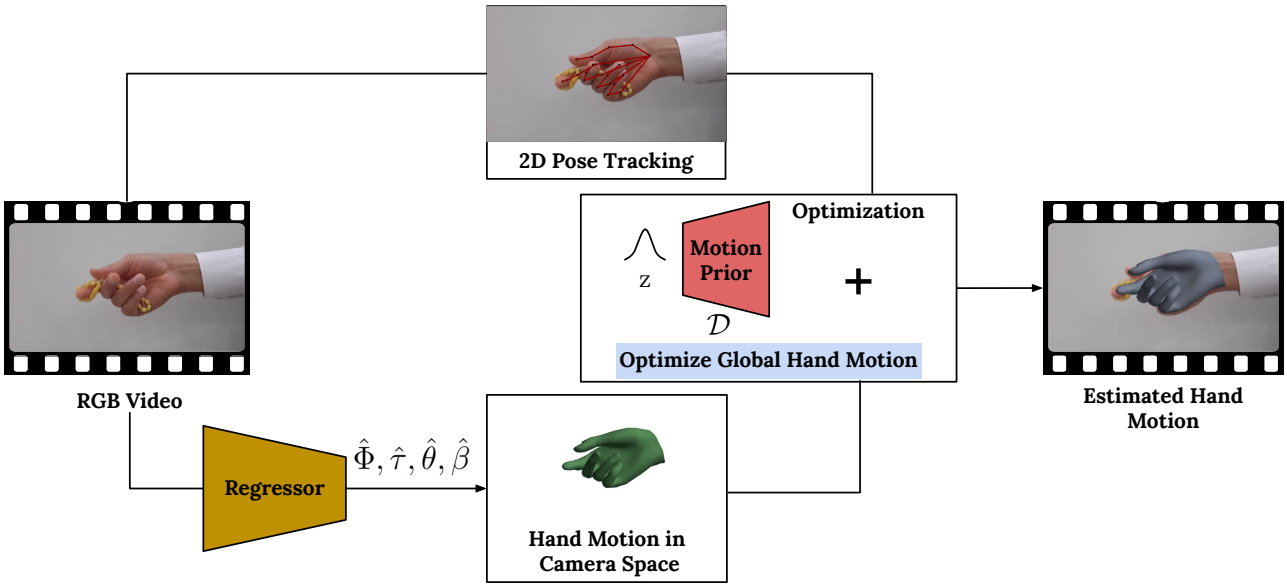

Figure 2. **HMP method overview:** Given a video of hand, we first use off-the-shelf methods to obtain initial 2D hand keypoints and MANO hand pose and shape parameters via the regressor. We propose a latent optimization framework that optimizes the hand motion to reduce 2D pose errors and increase motion likelihood under hand motion prior. The final output is temporally stable hand motion.

a detection for a timestep we project 3D joints obtained from PyMAF-X to the image space and use it as the keypoint source. We experimented with different keypoint estimation methods and empirically found that it is best to blend keypoints from Mediapipe and PyMAF-X (see Sec 4). This keypoint blending approach aims to combine strengths of each keypoint source for more accurate estimation. We refer to SupMat for more details.

3.3. Hand Motion Prior

Our objective is to build a motion prior to ensure the estimated hand motion's plausibility and to constrain the solution space during motion optimization. To achieve this, we employ a variational autoencoder (VAE) [31]. The VAE learns a latent representation, denoted as $\mathbf{z}$, of the hand motion and regularizes its latent code distribution to be a normal distribution. We want the decoder $\mathcal{D}$ of the VAE to be non-autoregressive for faster sampling while not sacrificing accuracy. Such a design choice becomes pivotal, especially when the motion prior is used iteratively during optimization. Autoregressive motion priors, such as HuMoR [45], tend to be unsuitably slow in handling long motion sequences. In contrast, a non-autoregressive decoder can be evaluated for the entire sequence in parallel. To this end, we adopt a Neural Motion Field (NeMF) [23] based decoder to represent body motion as a continuous vector field of hand poses via a NeRF-style MLP [39]. Section 3.4 discusses the application of NeMF for latent optimization.

Building on the approach from [23], our system solely models local hand motion using the prior. Specifically, $\mathcal{D}$ is an MLP accepting the latent code $\mathbf{z}_\theta$ and a time step t to produce the local hand pose $\hat{\theta}_t$ for the respective time step:

$$\mathcal{D} : (t, \mathbf{z}_\theta) \rightarrow (\hat{\theta}_t), \tag{1}$$

Here, $\mathbf{z}_\theta$ controls the local pose θ of the hand. Given a specific $\mathbf{z}_\theta$, the entire sequence can be sampled in parallel by simply varying the values of t. To incorporate the motion prior during latent optimization, we optimize the latent code $\{\mathbf{z}_\theta\}$ instead of solely optimizing the local hand motion $\{\theta_t\}_{t=0}^{T}$. We initialize the latent code using the pretrained encoder $\mathcal{E}$ of the VAE; *i.e.*, $z_\theta = \mathcal{E}_\theta(\{\theta\}_{t=0}^{T})$.

As opposed to to the approach in [23], we omit the global orientation information during motion prior training. This decision is grounded on the observation that hand global orientation is considerably less constrained. Unlike the body's global orientation, which is motion-limited by gravity, hands have the liberty to move freely in the air. Details for the training process are given in SupMat.

Stage	Variables	Loss terms	Description
1	Φ, τ, β	$\mathcal{L}_o, \mathcal{L}_{\text{tr}}, \mathcal{L}_\beta, \mathcal{L}_{os}, \mathcal{L}_{ts}, \mathcal{L}_{\text{2D}}$	global translation + rotation
2	$\Phi, \tau, \beta, \mathbf{z}_\theta$	$\mathcal{L}_o, \mathcal{L}_{\text{tr}}, \mathcal{L}_\beta, \mathcal{L}_{os}, \mathcal{L}_{\text{2D}}, \mathcal{L}_{MP}$	+ local hand pose

Table 1. Multi-stage optimization variables and loss terms.

3.4. Latent Optimization

This section explains the latent optimization process for estimating hand motion. Our goal is to optimize the vari-

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

Table 2. Multi-stage optimization loss coefficients.

ables expressing hand motion: global orientation Φ, global translation τ, shape β, and the motion prior latent code z_θ.

Objective Function: The objective function we aim to minimize is defined as:

$$\begin{aligned}\mathcal{L} = \lambda_o \mathcal{L}_o + \lambda_{\text{tr}} \mathcal{L}_{\text{tr}} + \lambda_\beta \mathcal{L}_\beta + \lambda_{os} \mathcal{L}_{os} + \lambda_{ts} \mathcal{L}_{ts} \\ + \lambda_{\text{MP}} \mathcal{L}_{\text{MP}} + \lambda_{\text{2D}} \mathcal{L}_{\text{2D}},\end{aligned} \tag{2}$$

where we use seven different objective terms with their corresponding coefficients. Values for these coefficients are given in Tab. 2. The first term ensures that the optimized global orientation is close to the initial global orientation through $\mathcal{L}_{\text{o}}$:

$$\mathcal{L}_{\text{o}} = \sum_{t=0}^{T} g(\Phi_t, \hat{\Phi}_t)^2. \tag{3}$$

where g is the geodesic distance between two rotations. The second term $\mathcal{L}_{\text{tr}}$ encourages the global translation not to diverge from its initial values:

$$\mathcal{L}_{\text{tr}} = \sum_{t=0}^{T} \|\tau_t - \hat{\tau}_t\|_2^2. \tag{4}$$

The third term motivates shape parameters to be close to zero vector through $\mathcal{L}_\beta$:

$$\mathcal{L}_\beta = \|\beta\|_2^2. \tag{5}$$

The fourth and fifth terms encourage smoothness of the global translation and orientation:

$$\mathcal{L}_{\text{os}} = \sum_{t=0}^{T-1} g(\Phi_{t+1}, \Phi_t)^2, \tag{6}$$

$$\mathcal{L}_{\text{ts}} = \sum_{t=0}^{T-1} \|\tau_{t+1} - \tau_t\|_2^2. \tag{7}$$

The 2D keypoint error term, $\mathcal{L}_{\text{2D}}$, constrains our motion to be aligned with the 2D keypoints predicted from detectors:

$$\mathcal{L}_{\text{2D}} = \sum_{i=1}^{21} \sum_{t \in T_{\text{detect}}} \alpha_t^i \rho \left(\Pi \left(R_{\text{cam}} J_t^i + T_{\text{cam}} \right) - \text{x}_t^i \right). \tag{8}$$

Here T_{detect} represents the time-steps where we have the keypoint detection. J_t^i stands for location of joint i in timestep t. x_t^i is the corresponding detected keypoint in image space, Π represents perspective projection to image space using camera intrinsics K, α_t^i represents the detection confidence for the joint i, and ρ is the Geman-McClure function [15]. The last term constrains hand motion to be valid by minimizing the negative log-likelihood of the latent code.

$$\mathcal{L}_{\text{MP}} = -\log \mathcal{N} \left(\mathbf{z}_\theta; \mu_\theta \left(\{\theta_t\} \right), \sigma_\theta \left(\{\theta_t\} \right) \right). \tag{9}$$

Multi-Stage Optimization: The estimation of 3D pose and shape from 2D video presents an inherently ill-posed problem. Attempting to optimize all parameters simultaneously can lead to local minima. To mitigate this, we adopt a gradual optimization process that progresses from coarse to fine-grained level. This approach serves the purpose of constraining the optimization problem at each stage.

Our optimization strategy unfolds in two distinct stages. During the initial stage, our objective is to align the initial hand estimations from PyMAF-X with the video data by optimizing the global orientation Φ, global translation τ, and shape β only. In the second stage, we include the motion prior latent code z to optimize the local pose, a phase that involves a more refined level of optimization. We use the Adam optimizer [30].

Occlusion Handling: For occluded frames, the off-the-shelf methods used for initialization, *i.e*. 2D keypoint and bounding-box detectors, PyMAF-X, do not provide any results. To robustly handle such frames without detection, we mask the objective terms in corresponding time steps, leaving us to solely optimize the latent code $\mathbf{z}_\theta$ with the observed time steps. Motion prior in such cases behave as an motion infilling method which infer the occluded frames with the cues from visible frames. This is a key part of our method which makes it robust to occlusions.

Parallel Optimization: A natural candidate for motion prior formulation is HuMoR [45]. However its autoregressive formulation renders it impractical for using long sequences. Instead, we aim to have a motion prior suitable for batch optimization. Our architecture is based on a recent work NeMF [23]. Its formulation allows parallel optimization, making it applicable to long motion sequences.

4. Experiments

4.1. Datasets and Metrics

HO3D is a dataset focused on capturing temporal interactions between hands and objects. This dataset comprises interactions of 10 subjects with 10 distinct YCB objects, all

Methods	HO3D-v3		
	PA-MPJPE ↓	RA-MPJPE ↓	RA-ACC ↓
TempCLR† [69]	10.6	-	3.7
HandOccNet† [41]	9.1	24.9	-
METRO [34]	12.1	38.7	17.4
PyMAF-X [64]	10.8	29.6	9.3
[34] + HMP (Ours)	10.8	31.3	2.4
[64] + HMP (Ours)	**10.1**	**26.7**	**2.2**

Table 3. State-of-the-art comparison on the HO3D-v3 dataset [18]. Methods denoted with † uses HO-3D as their training dataset.

Methods	DexYCB		
	PA-MPJPE ↓	RA-MPJPE ↓	RA-ACC ↓
ArtiBoost† [58]	-	12.8	-
Deformer† [14]	5.2	-	-
PyMAF-X [64]	11.6	38.1	17.1
[64] + HMP (Ours)	**8.9**	**34.1**	**3.6**

Table 4. State-of-the-art comparison on the DexYCB dataset [8]. Methods denoted with † uses DexYCB as their training dataset.

captured from various viewpoints [17, 55]. The manipulation of handheld objects within this dataset often results in substantial occlusions, posing challenges for analysis. We worked on version-3 of the dataset for evaluation.

HO3D-OCC: We choose occlusion-specific sequences from HO3D to highlight the performance of different methods under occlusion. This subset is derived from the training segment of HO3D and is comprised of 1736 frames.

DexYCB [8]: This dataset contains 10 subjects grasping 20 different objects from YCB-Video dataset [55]. Ground-truth values are obtained through an optimization process using hand-annotated 2D keypoints, and multiview RGB-D captures. The sequences are shorter (2-3 seconds) and motions have less articulation in comparison to HO3D [17]. The default split (**S0**) is used for evaluation.

AMASS [37] is a large dataset of 3D human motion capture curated from various marker-based datasets. Among many datasets included in the AMASS repository, GRAB [51], TCDHands, and SAMP [20] feature hand articulations. We use these datasets to train our motion prior.

Metrics: We report Procrustes aligned (PA-MPJPE), root aligned (RA-MPJPE) Mean-Per-Joint Projection Error in millimeters (mm). We also report root aligned acceleration error (RA-ACC) in mm/s^2. Acceleration error demonstrate the smoothness of estimated motion.

4.2. Comparison With the State-of-the-Art

Our aim is to have a method that generalize well to video from different sources. One way of ensuring that is to use a method that performs best in in-the-wild settings. PyMAF-X [64] is the current SOTA hand pose and shape estimation method. We use PyMAF-X as our main baseline and report other results based on it.

Our main goal in this paper is to recover coherent motion of hands. Therefore, we would like to emphasize metrics which measures the quality of the estimated motion *e.g.* RA-ACC. Unfortunately such metrics are not available for the evaluation server of the HO3D dataset. Therefore, in Tab. 3 we use the training set to report metrics and compare with methods doing so. All results listed on DexYCB Tab. 4 use **S0** subset.

Methods	HO3D-v3		
	PA-MPJPE ↓	RA-MPJPE ↓	RA-ACC ↓
PyMAF-X [64] + SLERP	10.7	29.4	5.9
No Motion Prior	10.5	28.0	1.9
PCA-based Prior	13.8	31.1	10.7
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 [64]	10.8	29.6	9.3
[64] + HMP (Ours)	**10.1**	**26.7**	2.2

Table 5. Ablation studies on the HO3D-v3 dataset [18].

Recent methods, such as Deformer [14] and ArtiBoost [58], use 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 a † in the corresponding tables. Overall, our method outperforms the existing state-of-the-art (SOTA) techniques on the HO3D and DexYCB datasets. Furthermore, our approach enhances the performance of the PyMAF-X method, which we employ for initialization, across both datasets.

Additionally, we provide qualitative results DexYCB in Fig. 3, on an in-the-wild video in Fig. 4, and on HO3D in Fig. 5. Please see the SupMat for more results. Compared to PyMAF-X method, our method is robust to occlusion caused by hand-object interaction.

To show the applicability of aforementioned plug-and-play fashion in Sec. 3.2, we also report quantitative numbers for optimization with different initialization methods in Tab. 3.

4.3. Ablation Study

We report the ablation experiments on HO3D and DexYCB datasets, in Tabs. 5 and 6 respectively. In this section, we analyze the critical components of our method.

Motion Prior: In addition to the NeMF-based motion prior we use, we report the results of using no motion prior, a

Methods	DexYCB PA-MPJPE ↓	RA-MPJPE ↓	RA-ACC ↓
PyMAF-X [64] + SLERP	11.5	36.5	6.0
No Motion Prior	10.9	36.4	3.4
PCA-based Prior	16.9	41.6	15.3
GMM-based Prior	10.8	38.7	4.8
Stage-1 (PyMAF-X)	10.8	35.1	**3.4**
Stage-1 (MediaPipe)	10.8	39.1	**3.4**
Stage-1 (MMPose)	10.8	39.4	**3.4**
Stage-1 (Blend)	10.8	35.5	**3.4**
Stage-2 (Blend)	**8.9**	**34.1**	3.6
PyMAF-X [64]	11.6	38.1	17.1
[64] + HMP (Ours)	**8.9**	**34.1**	3.6

Table 6. Ablation studies on the DexYCB dataset [8].

Methods	HO3D-OCC PA-MPJPE ↓	RA-MPJPE ↓	RA-ACC ↓
PyMAF-X [64]	15.3	48.9	26.0
PyMAF-X [64] + SLERP	14.4	41.3	7.9
Stage-1	13.0	38.2	**2.8**
Stage-2	**12.6**	**38.1**	3.0

Table 7. Performances on occlusion specific HO3D-OCC subset

PCA-based motion prior, and GMM-based motion prior, denoted as *No Motion Prior*, *PCA-based prior*, and *GMM-based prior* respectively. In the *No Motion Prior* experiments, we directly optimize the MANO hand pose instead of the latent code of the motion prior. We introduce a pose smoothness term to the pose optimization process, replacing the motion prior likelihood. Our motion prior trained on AMASS dataset outperforms these motion prior baselines on both the HO3D and DexYCB datasets.

Multi Stage: We run ablation studies to demonstrate the results of different stages of our optimization process. We show that Stage-2 which optimizes local pose through latent optimization help to improve results over the Stage-1.

Keypoint Blending: We analyzed the performance with different 2D hand keypoint detection algorithms: MMPose [9], MediaPipe [36], and PyMAF-X [64]. We also reported a variant where we blend keypoints from MediaPipe [36] and PyMAF-X [64]. We find out that blending MediaPipe and PyMAF-X are better 2D hand keypoint detectors compared to MMPose. Blending MediaPipe with the PyMAF-X give the best results overall.

5. Conclusion and Discussion

In this work we propose HMP, a latent optimization-based method for 3D hand pose and shape estimation from video. Motivated by the fact that existing video-based 3D hand datasets are insufficient for training feedforward models to generalize to in-the-wild scenarios, we develop a generative motion prior specific for hands, trained on the AMASS dataset and then employ motion this prior for video-based 3D hand motion estimation following a latent optimization approach.

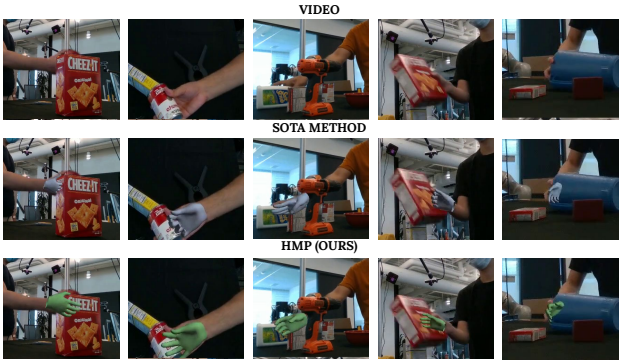

Figure 3. **3D hand pose and shape estimation on DexYCB videos:** input video (top), PyMAF-X (middle), HMP (bottom)

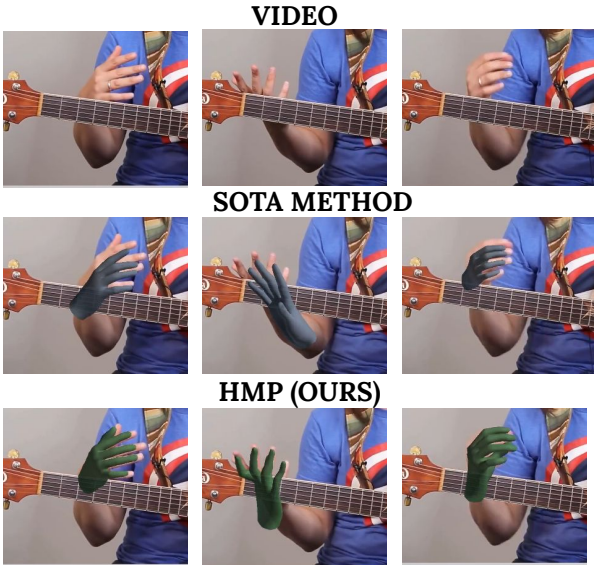

Figure 4. **3D hand pose and shape estimation on an in-the-wild video:** input video (top), PyMAF-X (middle), HMP (bottom)

Our integration of a robust motion prior significantly enhances performance, especially in occluded scenarios. It produces stable, temporally consistent results that surpass conventional single-frame methods. Our method's efficacy is demonstrated through both qualitative and quantitative evaluations on the HO3D and DexYCB datasets, with special emphasis on an occlusion-focused subset of HO3D. Our method can be used in plug-and-play fashion with any single-stage pose and shape regressor and improves its performance further. Due to this flexibility, unlike existing video hand pose and shape estimation methods, HMP works on in-the-wilds videos, too.

Limitations: A limitation of our approach is its reliance on 2D keypoint estimation quality. Existing 2D keypoint predictors can fail under heavy occlusion or motion blur.

Acknowledgements: We thank Nikos Athanasiou, Peter Kulits, and all Perceiving Systems department members for their valuable feedback and insightful discussions.

Disclosure: `https://files.is.tue.mpg.de/black/CoI_ICCV_2023.txt`

Figure 5. **3D hand pose and shape estimation on HO3D videos:** input video (top), PyMAF-X (middle), HMP (bottom)

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3D human motion modelling. In *ICCV*, 2019.

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering. In *CVPR*, pages 1067–1076, 2019.

[3] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: Probabilistic 3D human motion prediction via gan. In *CVPR Workshops*, 2018.

[4] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3D hand shape and pose from images in the wild. In *CVPR*, pages 10843–10852, 2019.

[5] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3D hand pose estimation from monocular RGB images. In *ECCV*, pages 666–682, 2018.

[6] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In *ICCV*, pages 2272–2281, 2019.

[7] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, pages 387–404. Springer, 2020.

[8] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *CVPR*, pages 9044–9053, 2021.

[9] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020.

[10] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *CVPR*, pages 6608–6617, 2020.

[11] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael J. Black, and Otmar Hilliges. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *3DV*, 2021.

[12] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023.

[13] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *ICCV*, 2015.

[14] 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.

[15] Stuart Geman and Donald E McClure. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, 1987.

[16] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. A neural temporal model for human motion prediction. In *CVPR*, 2019.

[17] 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.

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

[19] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

[20] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J. Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021.

[21] 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.

[22] 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*, pages 11807–11816, 2019.

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

[24] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *CVPR*, 2019.

[25] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3D scenes. In *CVPR*, 2023.

[26] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via 2.5D latent heatmap regression. In *ECCV*, 2018.

[27] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-RNN: Deep learning on spatio-temporal graphs. In *CVPR*, 2016.

[28] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *3DV*, 2020.

[29] Tarasha Khurana, Achal Dave, and Deva Ramanan. Detecting invisible people. In *ICCV*, pages 3174–3184, 2021.

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

[31] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[32] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020.

[33] Sijin Li and Antoni B Chan. 3D human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, pages 332–347. Springer, 2014.

[34] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021.

[35] 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.

[36] 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.

[37] 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.

[38] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017.

[39] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[40] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. GANerated hands for real-time 3D hand tracking from monocular RGB. In *CVPR*, 2018.

[41] 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.

[42] Sungheon Park, Jihye Hwang, and Nojun Kwak. 3D human pose estimation using convolutional neural networks with 2D pose information. In *ECCV*, pages 156–169. Springer, 2016.

[43] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. In *BMVC*, 2018.

[44] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV*, 2021.

[45] 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.

[46] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Siggraph Asia*, 36(6):245:1–245:17, Nov. 2017.

[47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

[48] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3D hand pose estimation via biomechanical constraints. In *ECCV*, pages 211–228. Springer, 2020.

[49] Adrian Spurr, Pavlo Molchanov, Umar Iqbal, Jan Kautz, and Otmar Hilliges. Adversarial motion modelling helps semi-supervised hand pose estimation. *arXiv preprint arXiv:2106.05954*, 2021.

[50] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, pages 89–98, 2018.

[51] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, pages 581–600, 2020.

[52] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In *CVPR*, pages 4511–4520, 2019.

[53] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023.

[54] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *ICML*, 2017.

[55] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. 2018.

[56] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. MT-VAE: Learning motion transformations to generate multimodal human dynamics. In *ECCV*, 2018.

[57] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. Seqhand: RGB-sequence-based 3D hand pose and shape estimation. In *ECCV*, pages 122–139. Springer, 2020.

[58] 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.

[59] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *CVPR*, pages 9877–9886, 2019.

[60] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, 2020.

[61] Ye Yuan and Kris Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. In *NeurIPS*, 2020.

[62] Ye Yuan and Kris M. Kitani. Diverse trajectory forecasting with determinantal point processes. In *ICLR*, 2020.

[63] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-guided human motion diffusion model. In *ICCV*, 2023.

[64] 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.

[65] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[66] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4D human body capture in 3D scenes. In *ICCV*, 2021.

[67] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular RGB image. In *ICCV*, pages 2354–2364, 2019.

[68] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *ICLR*, 2018.

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

[70] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, pages 4903–4911, 2017.