HuMoR: 3D Human Motion Model for Robust Pose Estimation

Davis Rempe[1] Tolga Birdal[1] Aaron Hertzmann[2] Jimei Yang[2]
Srinath Sridhar[3] Leonidas J. Guibas[1]
[1]Stanford University [2]Adobe Research [3]Brown University

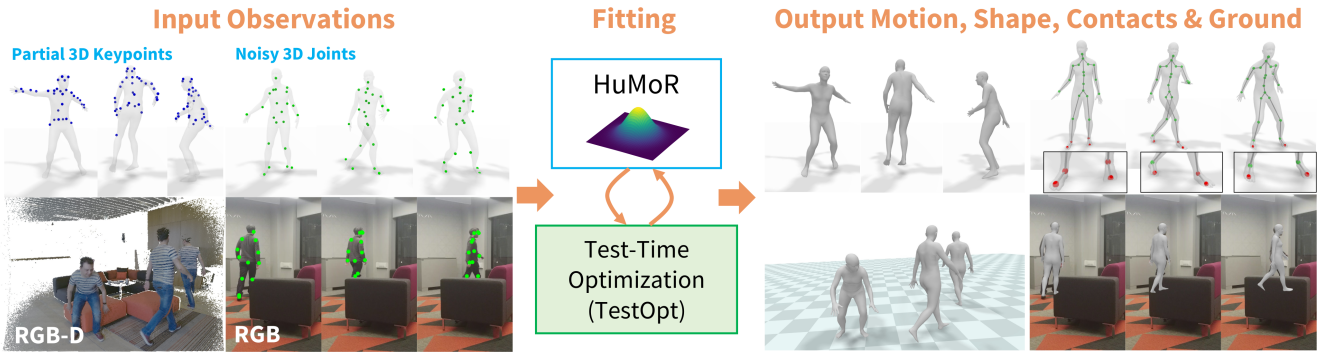

Figure 1: *Overview*. HuMoR is a 3D **Hu**man **Mo**tion model for **R**obust estimation of temporal pose formulated as a conditional variational autoencoder. (Left) The proposed approach can operate on many input modalities and is designed to handle partial and noisy observations. (Middle/Right) A test-time optimization fits 3D motion and shape to an input sequence using HuMoR as a prior; additional outputs include the ground and person-ground contacts (colored as ground plane and contacts).

Abstract

*We introduce HuMoR: a 3D **Hu**man **Mo**tion Model for **R**obust Estimation of temporal pose and shape. Though substantial progress has been made in estimating 3D human motion and shape from dynamic observations, recovering plausible pose sequences in the presence of noise and occlusions remains a challenge. For this purpose, we propose an expressive generative model in the form of a conditional variational autoencoder, which learns a distribution of the change in pose at each step of a motion sequence. Furthermore, we introduce a flexible optimization-based approach that leverages HuMoR as a motion prior to robustly estimate plausible pose and shape from ambiguous observations. Through extensive evaluations, we demonstrate that our model generalizes to diverse motions and body shapes after training on a large motion capture dataset, and enables motion reconstruction from multiple input modalities including 3D keypoints and RGB(-D) videos. See the project page at* `geometry.stanford.edu/projects/humor`.

1. Introduction

As humans, we are constantly moving in, interacting with, and manipulating the world around us. Thus, applications such as action recognition [79, 80] or holistic dynamic indoor scene understanding [15] require accurate perception of 3D human pose, shape, motion, contacts, and interaction. Extensive previous work has focused on estimating 2D or 3D human pose [13, 52, 53], shape [57, 26, 67], and motion [37] from videos. These are challenging problems due to the large space of articulations, body shape, and appearance variations. Even the best methods struggle to accurately capture a wide variety of motions from varying input modalities, producing noisy or overly-smoothed motions (especially at ground contact, *i.e.*, footskate), and struggle with occlusions (*e.g.*, walking behind a couch as in Fig. 1).

We focus on the problem of building a robust human motion model that can address these challenges. To date, most motion models directly represent sequences of likely poses — *e.g.*, in PCA space [55, 77, 70] or via future-predicting autoregressive processes [75, 76, 61]. However, purely pose-based predictions either make modeling environment interactions and generalization beyond training poses difficult, or quickly diverge from the space of realistic motions. On the other hand, explicit physical dynamics models [63, 43, 69, 62, 12, 11] are resource intensive and require knowledge of unobservable physical quantities. While generative models potentially offer the required flexibility, building an *expressive*, *generalizable* and *robust* model for *realistic* 3D human motions remains an open problem.

To address this, we introduce a learned, autoregressive, generative model that captures the *dynamics* of 3D human

motion, *i.e.*, how pose changes over time. Rather than describing likely poses, the **Hu**man **Mo**tion Model for **R**obust Estimation (HuMoR) models a probability distribution of possible *pose transitions*, formulated as a conditional variational autoencoder [72]. Though not explicitly physics-based, its components correspond to a physical model: the latent space can be interpreted as generalized forces, which are inputs to a dynamics model with numerical integration (the decoder). Moreover, ground contacts are explicitly predicted and used to constrain pose estimation at test time.

After training on the large AMASS motion capture dataset [51], we 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(-D) video and 2D or 3D joint sequences, as illustrated in Fig. 1 (left). In particular, we introduce a robust test-time optimization strategy which interacts with HuMoR to estimate the parameters of *3D motion*, *body shape*, the *ground plane*, and *contact points* as shown in Fig. 1 (middle/right). This interaction happens in two ways: (i) by parameterizing the motion in the latent space of HuMoR, and (ii) using HuMoR priors in order to regularize the optimization towards the space of plausible motions.

Comprehensive evaluations reveal that our method surpasses the state-of-the-art on a variety of visual inputs in terms of accuracy and physical plausibility of motions under partial and severe occlusions. We further demonstrate that our motion model generalizes to diverse motions and body shapes on common generative tasks like sampling and future prediction. In a nutshell, our contributions are:

- HuMoR, a generative 3D human motion prior modeled by a novel conditional VAE which enables expressive and general motion reconstruction and generation,
- A subsequent robust test-time optimization approach that uses HuMoR as a strong motion prior jointly solving for pose, body shape, and ground plane / contacts,
- The capability to operate on a variety of inputs, such as RGB(-D) video and 2D/3D joint position sequences, to yield accurate and plausible motions and contacts, exemplified through extensive evaluations.

Our work, more generally, suggests that neural nets for dynamics problems can benefit from architectures that model transitions, allowing control structures that emulate classical physical formulations.

2. Related Work

Much progress has been made on building methods to recover 3D joint locations [60, 53, 52] or parameterized 3D pose and shape (*i.e.*, SMPL [48]) from observations [78]. We focus primarily on motion and shape estimation.

Learning-Based Estimation. Deep learning approaches have shown success in regressing 3D shape and pose from a single image [39, 34, 58, 25, 24, 87, 16]. This has led to developments in predicting *motion* (pose sequences) and shape directly from RGB video [35, 89, 68, 74, 18]. Most recently, VIBE [37] uses adversarial training to encourage plausible outputs from a conditional recurrent motion generator. MEVA [50] maps a fixed-length image sequence to the latent space of a pre-trained motion autoencoder. These methods are fast and produce accurate root-relative joint positions for video, but motion is globally inconsistent and they struggle to generalize, *e.g.*, under severe occlusions. Other works have addressed occlusions but only on static images [7, 90, 64, 22, 38]. Our approach resolves difficult occlusions in video and other modalities by producing plausible and expressive motions with HuMoR.

Optimization-Based Estimation. One may directly optimize to more accurately fit to observations (images or 2D pose estimators [13]) using human body models [20, 4, 8]. SMPLify [8] uses the SMPL model [48] to fit pose and shape parameters to 2D keypoints in an image using priors on pose and shape. Later works consider body silhouettes [41] and use a learned variational pose prior [57]. Optimization for motion sequences has been explored by several works [3, 33, 47, 88, 83] which apply simple smoothness priors over time. These produce reasonable estimates when the person is fully visible, but with unrealistic dynamics, *e.g.*, overly smooth motions and footskate.

Some works employ human-environment interaction and contact constraints to improve shape and pose estimation [28, 47, 29] by assuming scene geometry is given. iMapper [54] recovers both 3D joints and a primitive scene representation from RGB video based on interactions by motion retrieval, which may differ from observations. In contrast, our approach optimizes for pose and shape by using an expressive generative model that produces more natural motions than prior work with realistic ground contact.

Human Motion Models. Early sophisticated motion models for pose tracking used a variety of approaches, including mixtures-of-Gaussians [32], linear embeddings of periodic motion [55, 77, 70], nonlinear embeddings [19], and nonlinear autoregressive models [75, 81, 76, 61]. These methods operate in pose space, and are limited to specific motions. Models based on physics can potentially generalize more accurately [63, 43, 69, 62, 12, 11, 86], while also estimating global pose and environmental interactions. However, general-purpose physics-based models are difficult to learn, computationally intensive at test-time, and often assume full-body visibility to detect contacts [63, 43, 69].

Many motion models have been learned for computer animation [10, 40, 66, 42, 46, 31, 73] including recent recurrent and autoregressive models [27, 23, 30, 84, 44]. These often focus on visual fidelity for a small set of characters

and periodic locomotions. Some have explored generating more general motion and body shapes [91, 59, 1, 17], but in the context of short-term future prediction. HuMoR is most similar to Motion VAE [44], however we make crucial contributions to enable generalization to unseen, non-periodic motions on novel body shapes.

3. HuMoR: 3D Human Dynamics Model

The goal of our work is to build an *expressive* and *generalizable* generative model of 3D human motion learned from real human motions, and to show that this can be used for robust test-time optimization (TestOpt) of pose and shape. In this section, we first describe the model, HuMoR.

State Representation. We represent the state of a moving person as a matrix $\mathbf{x}$ composed of a root translation $\mathbf{r} \in \mathbb{R}^3$, root orientation $\Phi \in \mathbb{R}^3$ in axis-angle form, body pose joint angles $\Theta \in \mathbb{R}^{3\times 21}$ and joint positions $\mathbf{J} \in \mathbb{R}^{3\times 22}$:

$$\mathbf{x} = [\quad \mathbf{r} \quad \dot{\mathbf{r}} \quad \Phi \quad \dot{\Phi} \quad \Theta \quad \mathbf{J} \quad \dot{\mathbf{J}} \quad], \tag{1}$$

where $\dot{\mathbf{r}}, \dot{\Phi}$ and $\dot{\mathbf{J}}$ denote the root and joint velocities, respectively, giving $\mathbf{x} \in \mathbb{R}^{3\times 69}$. Part of the state, $(\mathbf{r}, \Phi, \Theta)$, parameterizes the SMPL body model [48, 65] which is a differentiable function $M(\mathbf{r}, \Phi, \Theta, \beta)$ that maps to body mesh vertices $\mathbf{V} \in \mathbb{R}^{3\times 6890}$ and joints $\mathbf{J}^{\text{SMPL}} \in \mathbb{R}^{3\times 22}$ given shape parameters $\beta \in \mathbb{R}^{16}$. Our over-parameterization allows for two ways to recover the joints: (i) explicitly from $\mathbf{J}$, (ii) implicitly through the SMPL map $M(\cdot)$.

Latent Variable Dynamics Model. We are interested in modeling the probability of a time sequence of states

$$p_\theta(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T) = p_\theta(\mathbf{x}_0) \prod_{t=1}^{T} p_\theta(\mathbf{x}_t | \mathbf{x}_{t-1}), \tag{2}$$

where each state is assumed to be dependent on only the previous one and θ are learned parameters. Then $p_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})$ must capture the *plausibility* of a transition.

We propose a **conditional variational autoencoder (CVAE)** which formulates the motion $p_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})$ as a latent variable model as shown in Fig. 2. Following the original CVAE derivation [72], our model contains two main components. First, conditioned on the previous state $\mathbf{x}_{t-1}$, the distribution over possible latent variables $\mathbf{z}_t \in \mathbb{R}^{48}$ is described by a learned **conditional prior**:

$$p_\theta(\mathbf{z}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \mu_\theta(\mathbf{x}_{t-1}), \sigma_\theta(\mathbf{x}_{t-1})), \tag{3}$$

which parameterizes a Gaussian distribution with diagonal covariance via a neural network. Intuitively, the latent variable $\mathbf{z}_t$ represents the transition to $\mathbf{x}_t$ and should therefore have different distributions given different $\mathbf{x}_{t-1}$. For example, an idle person has a large variation of possible next states while a person in midair is on a nearly deterministic trajectory. Learning the conditional prior significantly improves the ability of the CVAE to generalize to diverse motions and empirically stabilizes both training and TestOpt.

Second, conditioned on $\mathbf{z}_t$ and $\mathbf{x}_{t-1}$, the **decoder** produces two outputs, Δ_θ and $\mathbf{c}_t$. The *change in state* Δ_θ defines the output distribution $p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{x}_{t-1})$ through

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \Delta_\theta(\mathbf{z}_t, \mathbf{x}_{t-1}) + \boldsymbol{\eta}, \qquad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4}$$

We find the additive update Δ_θ improves predictive accuracy compared to direct next-step prediction. The person-ground contact $\mathbf{c}_t$ is the probability that each of 8 body joints (*left and right toes*, *heels*, *knees*, and *hands*) is in contact with the ground at time t. Contacts are *not* part of the input to the conditional prior, only an output of the decoder. The contacts enable environmental constraints in TestOpt.

The complete probability model for a transition is then:

$$p_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}) = \int_{\mathbf{z}_t} p_\theta(\mathbf{z}_t|\mathbf{x}_{t-1}) p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{x}_{t-1}). \tag{5}$$

Given an initial state $\mathbf{x}_0$, one can sample a motion sequence by alternating between sampling $\mathbf{z}_t \sim p_\theta(\mathbf{z}_t|\mathbf{x}_{t-1})$ and sampling $\mathbf{x}_t \sim p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{x}_{t-1})$, from $t = 1$ to T. This model parallels a conventional stochastic physical model. The conditional prior can be seen as a controller, producing "forces" $\mathbf{z}_t$ as a function of state $\mathbf{x}_{t-1}$, while the decoder acts like a combined physical dynamics model and Euler integrator of generalized position and velocity in Eq. (4).

In addition to this nice physical interpretation, our model is motivated by Motion VAE (MVAE) [44], which has recently shown promising results for single-character locomotion animation, also using a VAE for $p_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})$. However, we find that directly applying MVAE for estimation does not give good results (Sec. 5). We overcome this by additionally learning a *conditional* prior, modeling the *change in state* and contacts, and encouraging *consistency* between joint position and angle predictions (Sec. 3.1).

Rollout. We use our model to define a deterministic *rollout function*, which is key to TestOpt. Given an initial state $\mathbf{x}_0$ and a sequence of latent transitions $\mathbf{z}_{1:T}$, we define a function $\mathbf{x}_T = f(\mathbf{x}_0, \mathbf{z}_{1:T})$ that deterministically maps the motion "parameters" $(\mathbf{x}_0, \mathbf{z}_{1:T})$ to the resulting state at time T. This is done through autoregressive rollout which decodes and integrates $\mathbf{x}_t = \mathbf{x}_{t-1} + \Delta_\theta(\mathbf{z}_t, \mathbf{x}_{t-1})$ at each timestep.

Initial State GMM. We model $p_\theta(\mathbf{x}_0)$ with a Gaussian mixture model (GMM) containing $K = 12$ components with weights γ^i, so that $p_\theta(\mathbf{x}_0) = \sum_{i=1}^{K} \gamma^i \mathcal{N}(\mathbf{x}_0; \mu_\theta^i, \sigma_\theta^i)$.

3.1. Training

Our CVAE is trained using pairs of $(\mathbf{x}_{t-1}, \mathbf{x}_t)$. We consider the usual variational lower bound:

$$\log p_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}) \geq \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{x}_{t-1})] - D_{\text{KL}}(q_\phi(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t-1}) \,\|\, p_\theta(\mathbf{z}_t|\mathbf{x}_{t-1})). \tag{6}$$

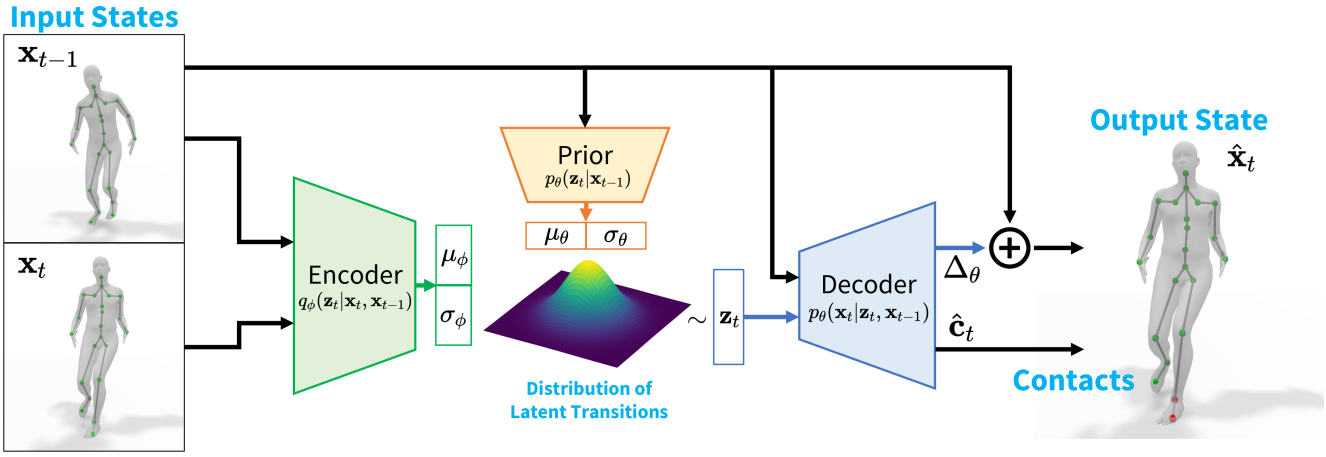

Figure 2: *HuMoR CVAE Architecture*. During training, given the previous state $\mathbf{x}_{t-1}$ and ground truth current state $\mathbf{x}_t$, the model reconstructs $\hat{\mathbf{x}}_t$ by sampling from the **encoder** distribution. At test time we can (i) *generate* the next state from $\mathbf{x}_{t-1}$ by sampling from the **prior** distribution and **decoding**, (ii) *infer* a latent transition $\mathbf{z}_t$ with the **encoder**, or (iii) evaluate the *likelihood* of a given $\mathbf{z}_t$ with the **conditional prior**.

The expectation term measures the reconstruction error of the **decoder**. The **encoder**, *i.e.* approximate posterior, is introduced for training and parameterizes a Gaussian distribution $q_\phi(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \mu_\phi(\mathbf{x}_t, \mathbf{x}_{t-1}), \sigma_\phi(\mathbf{x}_t, \mathbf{x}_{t-1}))$. The KL divergence $D_{\text{KL}}(\cdot \parallel \cdot)$ regularizes its output to be near the **prior**. Therefore, we seek the parameters (θ, ϕ) that minimize the loss function

$$\mathcal{L}_{\text{rec}} + w_{\text{KL}}\mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{reg}} \tag{7}$$

over all training pairs in our dataset, where $\mathcal{L}_{\text{rec}} + w_{\text{KL}}\mathcal{L}_{\text{KL}}$ is the lower bound in Eq. (6) with weight w_{KL}, and $\mathcal{L}_{\text{reg}}$ contains additional regularizers.

For a single training pair $(\mathbf{x}_{t-1}, \mathbf{x}_t)$, the reconstruction loss is computed as $\mathcal{L}_{\text{rec}} = ||\mathbf{x}_t - \hat{\mathbf{x}}_t||^2$ from the decoder output $\hat{\mathbf{x}}_t = \mathbf{x}_{t-1} + \Delta_\theta(\mathbf{z}_t, \mathbf{x}_{t-1})$ with $\mathbf{z}_t \sim q_\phi(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t-1})$. Gradients are backpropagated through this sample using the reparameterization trick [36]. The regularization loss contains two terms: $\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{SMPL}} + w_{\text{contact}}\mathcal{L}_{\text{contact}}$. The SMPL term $\mathcal{L}_{\text{SMPL}} = \mathcal{L}_{\text{joint}} + \mathcal{L}_{\text{vtx}} + \mathcal{L}_{\text{consist}}$ uses the output of the body model with the estimated parameters and ground truth shape $[\hat{\mathbf{J}}_t^{\text{SMPL}}, \hat{\mathbf{V}}_t] = M(\hat{\mathbf{r}}_t, \hat{\Phi}_t, \hat{\Theta}_t, \beta)$:

$$\mathcal{L}_{\text{joint}} = ||\mathbf{J}_t^{\text{SMPL}} - \hat{\mathbf{J}}_t^{\text{SMPL}}||^2 \tag{8}$$

$$\mathcal{L}_{\text{vtx}} = ||\mathbf{V}_t - \hat{\mathbf{V}}_t||^2 \qquad \mathcal{L}_{\text{consist}} = ||\hat{\mathbf{J}}_t - \hat{\mathbf{J}}_t^{\text{SMPL}}||^2. \tag{9}$$

The loss $\mathcal{L}_{\text{consist}}$ encourages consistency between regressed joints and those of the body model. The contact loss $\mathcal{L}_{\text{contact}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{vel}}$ contains two terms. The first supervises ground contact classification with a typical binary cross entropy; the second regularizes joint velocities to be consistent with contacts $\mathcal{L}_{\text{vel}} = \sum_j \hat{c}_t^j ||\hat{\mathbf{v}}_t||^2$ with $\hat{\mathbf{v}}_t \in \hat{\dot{\mathbf{J}}}_t$ and $\hat{c}_t^j \in \hat{\mathbf{c}}_t$ the predicted probability that joint j is in ground contact. We set $w_{\text{contact}} = 0.01$ and $w_{\text{KL}} = 4e^{-4}$.

The initial state GMM is trained separately with expectation-maximization on the same dataset used to train the CVAE.

Implementation Details. To ease learning and improve generalization, our model operates in an aligned canonical coordinate frame at each step. All networks are 4 or 5 layer MLPs with ReLU activations and group normalization [82]. To combat posterior collapse [49, 44, 72], we linearly anneal w_{KL} during training [9]. Following [44], we also use scheduled sampling [6] to enable long-term generation by making the model robust to its own errors. Additional details are available in the supplementary.

4. Test-time Motion Optimization

We next use the space of motion learned by HuMoR as a *prior* in TestOpt to recover pose and shape from noisy and partial observations while ensuring plausibility.

4.1. Optimization Variables

Given a sequence of observations $\mathbf{y}_{0:T}$, either as 2D/3D joints, 3D point clouds, or 3D keypoints, we seek the shape β and a sequence of SMPL pose parameters $(\mathbf{r}_{0:T}, \Phi_{0:T}, \Theta_{0:T})$ which describe the underlying motion being observed. We parameterize the optimized motion **using our CVAE** by the initial state $\mathbf{x}_0$ and a sequence of latent transitions $\mathbf{z}_{1:T}$. Then at T (and any intermediate steps) $\mathbf{x}_T = f(\mathbf{x}_0, \mathbf{z}_{1:T})$ is determined through model *rollout* using the **decoder** as previously detailed. Compared to directly optimizing SMPL [3, 8, 33], this motion representation naturally encourages plausibility and is compact in the number of variables. To obtain the transformation between the canonical coordinate frame in which our CVAE is trained and the observation frame used for optimization, we

additionally optimize the ground plane of the scene $\mathbf{g} \in \mathbb{R}^3$. All together, we simultaneously optimize initial state $\mathbf{x}_0$, a sequence of latent variables $\mathbf{z}_{1:T}$, ground $\mathbf{g}$, and shape β. We assume a static camera with known intrinsics.

4.2. Objective & Optimization

The optimization objective can be formulated as a maximum a-posteriori (MAP) estimate (see supplementary), which seeks a motion that is plausible under our generative model while closely matching observations:

$$\min_{\mathbf{x}_0, \mathbf{z}_{1:T}, \mathbf{g}, \beta} \mathcal{E}_{\text{mot}} + \mathcal{E}_{\text{data}} + \mathcal{E}_{\text{reg}}. \tag{10}$$

We next detail each of these terms which are the motion prior, data, and regularization energies. In the following, λ are weights to determine the contribution of each term.

Motion Prior $\mathcal{E}_{\text{mot}}$. This energy measures the likelihood of the latent transitions $\mathbf{z}_{1:T}$ and initial state $\mathbf{x}_0$ under the HuMoR CVAE and GMM. It is $\mathcal{E}_{\text{mot}} = \mathcal{E}_{\text{CVAE}} + \mathcal{E}_{\text{init}}$ where

$$\begin{aligned} \mathcal{E}_{\text{CVAE}} &= -\lambda_{\text{CVAE}} \sum_{t=1}^{T} \log \mathcal{N}(\mathbf{z}_t; \mu_\theta(\mathbf{x}_{t-1}), \sigma_\theta(\mathbf{x}_{t-1})) \\ \mathcal{E}_{\text{init}} &= -\lambda_{\text{init}} \log \sum_{i=1}^{K} \gamma^i \mathcal{N}(\mathbf{x}_0; \mu_\theta^i, \sigma_\theta^i). \end{aligned} \tag{11}$$

$\mathcal{E}_{\text{CVAE}}$ uses the learned **conditional prior** and $\mathcal{E}_{\text{init}}$ uses the initial state GMM.

Data Term $\mathcal{E}_{\text{data}}$. This term is the *only* modality-dependent component of our approach, requiring different losses for different inputs: 3D joints, 2D joints, and 3D point clouds. All data losses operate on SMPL joints or mesh vertices obtained through the body model $[\mathbf{J}_t^{\text{SMPL}}, \mathbf{V}_t] = M(\mathbf{r}_t, \Phi_t, \Theta_t, \beta)$ using the current shape β along with the SMPL parameters $(\mathbf{r}_t, \Phi_t, \Theta_t)$ contained in $\mathbf{x}_t = f(\mathbf{x}_0, \mathbf{z}_{1:t})$ transformed from the canonical to observation (*i.e.* camera) frame. In the simplest case, the observations $\mathbf{y}_t$ are 3D joint positions (or keypoints with known correspondences) and our energy is

$$\mathcal{E}_{\text{data}} \triangleq \mathcal{E}_{\text{data}}^{\text{3D}} = \lambda_{\text{data}} \sum_{t=0}^{T} \sum_{j=1}^{J} ||\mathbf{p}_t^j - \mathbf{y}_t^j||^2 \tag{12}$$

with $\mathbf{p}_t^j \in \mathbf{J}_t^{\text{SMPL}}$. For 2D joint positions, each with a detection confidence σ_t^j, we use a re-projection loss

$$\mathcal{E}_{\text{data}} \triangleq \mathcal{E}_{\text{data}}^{\text{2D}} = \lambda_{\text{data}} \sum_{t=0}^{T} \sum_{j=1}^{J} \sigma_t^j \rho(\Pi(\mathbf{p}_t^j) - \mathbf{y}_t^j) \tag{13}$$

with ρ the robust Geman-McClure function [8, 21] and Π the pinhole projection. If an estimated person segmentation mask is available, it is used to ignore spurious 2D joints. Finally, if $\mathbf{y}_t$ is a 3D point cloud obtained from a depth map roughly masked around the person of interest, we use the mesh vertices to compute

$$\mathcal{E}_{\text{data}} \triangleq \mathcal{E}_{\text{data}}^{\text{PC3D}} = \lambda_{\text{data}} \sum_{t=0}^{T} \sum_{i=1}^{N_t} w_{\text{bs}} \min_{\mathbf{p}_t \in \mathbf{V}_t} ||\mathbf{p}_t - \mathbf{y}_t^i||^2 \tag{14}$$

where w_{bs} is a robust bisquare weight [5] computed based on the Chamfer distance term.

Regularizers $\mathcal{E}_{\text{reg}}$. The additional regularization consists of four terms $\mathcal{E}_{\text{reg}} = \mathcal{E}_{\text{skel}} + \mathcal{E}_{\text{env}} + \mathcal{E}_{\text{gnd}} + \mathcal{E}_{\text{shape}}$. The first two terms encourage rolled-out motions from the CVAE to be plausible even when the initial state $\mathbf{x}_0$ is far from the optimum (*i.e.* early in optimization). The skeleton consistency term uses the joints $\mathbf{J}_t$ directly *predicted* by the decoder during rollout along with the SMPL joints:

$$\mathcal{E}_{\text{skel}} = \sum_{t=1}^{T} \left(\lambda_{\text{c}} \sum_{j=1}^{J} ||\mathbf{p}_t^j - \mathbf{p}_t^{j,\text{pred}}||^2 + \lambda_{\text{b}} \sum_{i=1}^{B} (l_t^i - l_{t-1}^i)^2 \right)$$

with $\mathbf{p}_t^j \in \mathbf{J}_t^{\text{SMPL}}$ and $\mathbf{p}_t^{j,\text{pred}} \in \mathbf{J}_t$. The second summation uses bone lengths l computed from $\mathbf{J}_t$ at each step. The second regularizer $\mathcal{E}_{\text{env}}$ ensures consistency between predicted CVAE contacts, the motion, and the environment:

$$\mathcal{E}_{\text{env}} = \sum_{t=1}^{T} \sum_{j=1}^{J} \lambda_{\text{cv}} c_t^j ||\mathbf{p}_t^j - \mathbf{p}_{t-1}^j||^2 + \lambda_{\text{ch}} c_t^j \max(|\mathbf{p}_{z,t}^j| - \delta, 0)$$

where $\mathbf{p}_t^j \in \mathbf{J}_t^{\text{SMPL}}$ and c_t^j is the contact probability output from the model for joint j. The contact height term weighted by λ_{ch} ensures the z-component of contacting joints are within δ of the floor in the canonical frame.

The final two regularizers are priors on the ground and shape. We assume the ground should stay close to initialization $\mathcal{E}_{\text{gnd}} = \lambda_{\text{gnd}} ||\mathbf{g} - \mathbf{g}^{\text{init}}||^2$. Finally, β should stay near the neutral zero vector similar to [28, 57]: $\mathcal{E}_{\text{shape}} = \lambda_{\text{shape}} ||\beta||^2$.

Initialization & Optimization. We initialize the temporal SMPL parameters $\mathbf{r}_{0:T}, \Phi_{0:T}, \Theta_{0:T}$ and shape β with an initialization optimization using $\mathcal{E}_{\text{data}}$ and $\mathcal{E}_{\text{shape}}$ along with two additional regularization terms. $\mathcal{E}_{\text{pose}} = \sum_t ||\mathbf{z}_t^{\text{pose}}||^2$ is a pose prior where $\mathbf{z}_t^{\text{pose}} \in \mathbb{R}^{32}$ is the body joint angles represented in the latent space of the VPoser model [57, 28]. The smoothness term $\mathcal{E}_{\text{smooth}} = \sum_{t=1}^{T} \sum_{j=1}^{J} ||\mathbf{p}_t^j - \mathbf{p}_{t-1}^j||^2$ with $\mathbf{p}_t^j \in \mathbf{J}_t^{\text{SMPL}}$ smooths 3D joint positions over time. Afterwards, the initial latent sequence $\mathbf{z}_{1:T}^{\text{init}}$ is computed through inference with the CVAE **encoder**. Our optimization is implemented in PyTorch [56] using L-BFGS and *autograd*; with batching, a $3s$ RGB video takes about 5.5 min to fit. We provide further details in the supplementary material.

5. Experimental Results

We evaluate HuMoR on (i) generative sampling tasks and (ii) as a prior in TestOpt to estimate motion from 3D

Model	Future Prediction			Diversity
	Contact ↑	**ADE** ↓	**FDE** ↓	**APD** ↑
MVAE [44]	-	25.8	50.6	85.4
HuMoR	0.88	**21.5**	**42.1**	94.9
HuMoR (Qual)	0.88	22.0	46.3	**100.0**

Table 1: (Left) Future prediction accuracy for $2s$ AMASS sequences. Contact classification accuracy, average displacement error (cm), and final displacement error (cm) are reported. (Right) Sampling diversity over $5s$ rollouts measured by average pairwise distance (cm).

and RGB(-D) inputs. We encourage viewing the **supplementary video** to appreciate the qualitative improvement of our approach. Additional dataset and experiment details are available in the supplementary document.

5.1. Datasets

AMASS [51] is a large motion capture database containing diverse motions and body shapes on the SMPL body model. We sub-sample the dataset to 30 Hz and use the recommended training split to train the CVAE and initial state GMM in HuMoR. We evaluate on the held out Transitions and HumanEva [71] subsets (Sec. 5.3 and 5.4).

i3DB [54] contains RGB videos of person-scene interactions involving medium to heavy occlusions. It provides annotated 3D joint positions and a primitive 3D scene reconstruction which we use to fit a ground plane for computing plausibility metrics. We run off-the-shelf 2D pose estimation [13], person segmentation [14], and plane detection [45] models to obtain inputs for our optimization.

PROX [28] contains RGB-D videos of people interacting with indoor environments. We use a subset of the qualitative data to evaluate plausibility metrics using a floor plane fit to the provided ground truth scene mesh. We obtain 2D pose, person masks, and ground plane initialization in the same way as done for i3DB.

5.2. Baselines and Evaluation Metrics

Motion Prior Baselines. We ablate the proposed CVAE to analyze its core components: *No Delta* directly predicts the next state from the decoder rather than the change in state, *No Contacts* does not classify ground contacts, *No* $\mathcal{L}_{SMPL}$ does not use SMPL regularization in training, and *Standard Prior* uses $\mathcal{N}(\mathbf{0}, \mathbf{I})$ rather than our learned **conditional prior**. All of these ablated together recovers *MVAE* [44].

Motion Estimation Baselines. *VPoser-t* is the initialization phase of our optimization. It uses VPoser [57] and 3D joint smoothing similar to previous works [3, 33, 88]. *PROX-(RGB/D)* [28] are optimization-based methods which operate on individual frames of RGB and RGB-D videos, respectively. Both assume the full scene mesh is given to enforce contact and penetration constraints. *VIBE* [37] is a recent learned method to recover shape and pose from video.

Error Metrics. 3D positional errors are measured on joints, keypoints, or mesh vertices (**Vtx**) and compute *global* mean per-point position error unless otherwise specified. We report positional errors for all (**All**), occluded (**Occ**), and visible (**Vis**) observations separately. Finally, we report binary classification accuracy of the 8 person-ground contacts (**Contact**) predicted by HuMoR.

Plausibility Metrics. We use additional metrics to measure qualitative motion characteristics that joint errors cannot capture. Smoothness is evaluated by mean per-joint accelerations (**Accel**) [35]. Another important indicator of plausibility is ground penetration [63]. We use the true ground plane to compute the frequency (**Freq**) of foot-floor penetrations: the fraction of frames for both the left and right toe joints that penetrate more than a threshold. We measure frequency at 0, 3, 6, 9, 12, and 15 cm thresholds and report the mean. We also report mean penetration distance (**Dist**), where non-penetrating frames contribute a distance of 0 to make values comparable across differing frequencies.

5.3. Generative Model Evaluation

We first evaluate HuMoR as a standalone generative model and show improved generalization to unseen motions and bodies compared to *MVAE* for two common tasks (see Table 1): future prediction and diverse sampling. We use $2s$ AMASS sequences and start generation from the first step. Results are shown for *HuMoR* and a modified *HuMoR (Qual)* that uses $\mathbf{J}^{\text{SMPL}}$ as input to each step during rollout instead of $\mathbf{J}$, thereby enforcing skeleton consistency. This version produces *qualitatively* superior results for generation, but is too expensive to use during TestOpt.

For prediction, we report average displacement error (**ADE**) and final displacement error (**FDE**) [85], which measure mean joint errors over all steps and at the final step, respectively. We sample 50 $2s$ motions for each initial state and the one with lowest **ADE** is considered the prediction. For diversity, we sample 50 $5s$ motions and compute the average pairwise distance (**APD**) [2], *i.e.* the mean joint distance between all pairs of samples.

As seen in Tab. 1, the base MVAE [44] does not generalize well when trained on the large AMASS dataset; our proposed CVAE improves both the accuracy and diversity of samples. *HuMoR (Qual)* hinders prediction accuracy, but gives better diversity and visual quality (see supplement).

5.4. Estimation from 3D Observations

Next, we show that HuMoR also generalizes better when used in TestOpt for fitting to 3D data, and that using a motion prior is crucial to plausibly handling occlusions. $3s$ AMASS sequences are used to demonstrate key abilities: (i) fitting to partial data and (ii) denoising. For the former,

Method	Input	Positional Error Vis	Occ	All	Joints Legs	Mesh Vtx	Contact	Ground Pen Accel	Freq	Dist
VPoser-t	Occ Keypoints	**0.67**	20.76	9.22	21.08	7.95	-	5.71	16.77%	2.28
MVAE [44]	Occ Keypoints	2.39	19.15	9.52	16.86	8.90	-	7.12	**3.15%**	0.30
HuMoR (Ours)	Occ Keypoints	1.46	**17.40**	**8.24**	**15.42**	**7.56**	0.89	**5.38**	3.31%	**0.26**
VPoser-t	Noisy Joints	-	-	3.67	4.47	4.98	-	**4.61**	1.35%	0.07
MVAE [44]	Noisy Joints	-	-	2.68	3.21	4.42	-	6.5	1.75%	0.11
HuMoR (Ours)	Noisy Joints	-	-	**2.27**	**2.61**	**3.55**	0.97	5.23	**1.18%**	**0.05**

Table 2: Motion and shape estimation from 3D observations: partially occluded keypoints (top) and noisy joints (bottom). *Positional Error* (cm) is reported w.r.t. the input modality. Acceleration is m/s^2 and penetration distance in cm.

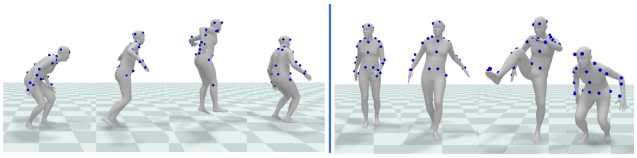

Figure 3: Fitting to partial 3D keypoints. HuMoR captures non-periodic motions like jumping, crouching, and kicking.

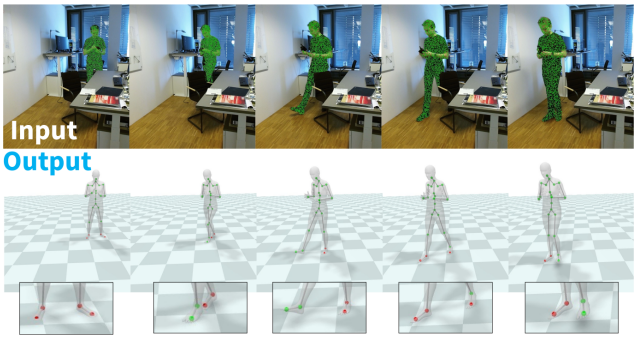

Figure 4: From RGB-D (top) TestOpt with HuMoR outputs 3D motion, the ground plane, and contacts (bottom).

TestOpt fits to 43 keypoints on the body that resemble motion capture markers; keypoints that fall below $0.9m$ at each timestep are "occluded", leaving the legs unobservable at most steps. For denoising, Gaussian noise with $4cm$ standard deviation is added to 3D joint position observations.

Tab. 2 compares to *VPoser-t* and to using *MVAE* as the motion prior during optimization rather than HuMoR. We report leg joint errors (toes, ankles, and knees), which are often occluded, separately. The right side of the table reports plausibility metrics. HuMoR gives more accurate poses, especially for occluded keypoints and leg joints. It also estimates smoother motions with fewer and less severe ground penetrations. For denoising, *VPoser-t* oversmooths which gives the lowest acceleration but least accurate motion. TestOpt with HuMoR gives inherently smooth results while still allowing for necessarily large accelerations to fit dynamic observations. Notably, HuMoR predicts person-ground contact with 97% accuracy even under severe noise. Qualitative results are shown in Fig. 1 and Fig. 3.

5.5. Estimation from RGB(-D) Observations

Finally, we show that TestOpt with HuMoR can be applied to real-world RGB and RGB-D observations, and outperforms baselines on positional and plausibility metrics especially from partial and noisy data. We use $3s$ (90 frame) clips from i3DB [54] and PROX [28]. Tab. 3 shows results on i3DB which affords quantitative 3D joint evaluation. The top half compares to baseline estimation methods; the bottom uses ablations of HuMoR in TestOpt rather than the full model. Mean per-joint position errors are reported for **global** joint positions and after root alignment.

As seen in Tab. 3, *VIBE* gives locally accurate predictions for visible joints, but large global errors and unrealistic accelerations due to occlusions and temporal inconsistency (see Fig. 5). *VPoser-t* gives reasonable global errors, but suffers frequent penetrations as shown for sitting in Fig. 5. Using *MVAE* or ablations of HuMoR as the motion prior in TestOpt fails to effectively generalize to real-world data and performs worse than the full model. The conditional prior and $\mathcal{L}_{\text{SMPL}}$ have the largest impact, while performance even without using contacts still outperforms the baselines.

The top half of Tab. 4 evaluates plausibility on additional RGB results from PROX compared to *VIBE* and *PROX-RGB*. Since *PROX-RGB* uses the scene mesh as input to enforce environment constraints, it is a very strong baseline and its performance on penetration metrics is expectedly good. HuMoR comparatively increases penetration frequency since it only gets a rough ground plane as initialization, but gives much smoother motions.

The bottom half of Tab. 4 shows results fitting to RGB-D for the same PROX data, which uses both $\mathcal{E}_{\text{data}}^{\text{2D}}$ *and* $\mathcal{E}_{\text{data}}^{\text{PC3D}}$ in TestOpt. This improves performance using HuMoR, slightly outperforming *PROX-D* which is less robust to issues with 2D joint detections and 3D point noise causing large errors. Qualitative examples are in Fig. 1 and Fig. 4.

Thanks to the generalizability of HuMoR, TestOpt is also effective in recovering very dynamic motions like dancing from RGB video when the full body is visible (see supplementary material for examples).

Method	Global Joint Error				Root-Aligned Joint Error				Ground Pen		
	Vis	Occ	All	Legs	Vis	Occ	All	Legs	Accel	Freq	Dist
VIBE [37]	90.05	192.55	116.46	121.61	12.06	23.78	15.08	21.65	243.36	7.98%	3.01
VPoser-t	28.33	40.97	31.59	35.06	12.77	26.48	16.31	25.60	4.46	9.28%	2.42
MVAE [44]	37.54	50.63	40.91	44.42	16.00	28.32	19.17	26.63	4.96	7.43%	1.55
No Delta	27.55	35.59	29.62	32.14	**11.92**	23.10	14.80	21.65	3.05	2.84%	**0.58**
No Contacts	26.65	39.21	29.89	35.73	12.24	23.36	15.11	22.25	**2.43**	5.59%	1.70
No $\mathcal{L}_{\text{SMPL}}$	31.09	43.67	34.33	36.84	12.81	25.47	16.07	23.54	3.21	4.12%	1.31
Standard Prior	77.60	146.76	95.42	99.01	18.67	39.40	24.01	34.02	5.98	8.30%	6.47
HuMoR (Ours)	**26.00**	**34.36**	**28.15**	**31.26**	12.02	**21.70**	**14.51**	**20.74**	**2.43**	**2.12%**	0.68

Table 3: Motion and shape from RGB video (*i.e.* 2D joints) on i3DB [54]. Joint errors are in cm and acceleration is m/s^2. Top shows results from motion estimation baselines while bottom uses ablations of HuMoR during optimization.

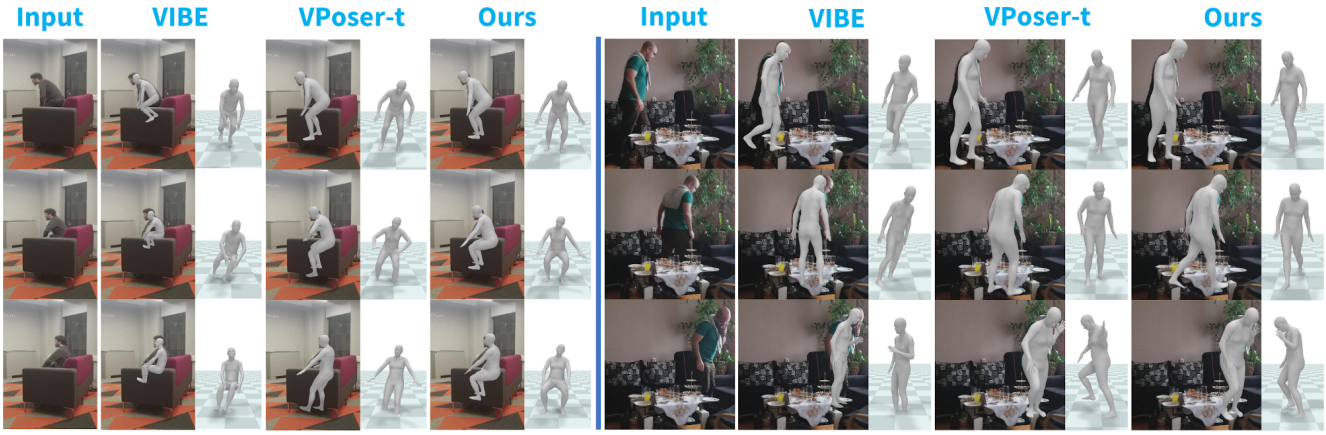

Figure 5: Qualitative comparison for fitting to RGB video (*i.e.* 2D joints) from i3DB [54]. Optimization using HuMoR (Ours) outputs natural and plausible sitting and walking motions under heavy occlusions compared to baseline approaches.

Method	Input	Ground Pen		
		Accel	Freq	Dist
VIBE [37]	RGB	86.06	23.46%	4.71
PROX-RGB [28]	RGB	196.07	**2.55%**	**0.32**
VPoser-t	RGB	3.14	13.38%	2.82
HuMoR (Ours)	RGB	**1.73**	9.99%	1.56
PROX-D [28]	RGB-D	46.59	8.95%	1.19
VPoser-t	RGB-D	3.27	10.66%	2.18
HuMoR (Ours)	RGB-D	**1.61**	**5.19%**	**0.85**

Table 4: Plausibility evaluation on videos in PROX [28]. Acceleration is m/s^2 and penetration distance in cm.

6. Discussion

We have introduced HuMoR, a learned generative model of 3D human motion leveraged during test-time optimization to robustly recover pose and shape from 3D, RGB, and RGB-D observations. We have demonstrated that the key components of our model enable generalization to novel motions and body shapes for both generative tasks and downstream optimization. Compared to strong learning and optimization-based baselines, HuMoR excels at estimating plausible motion under heavy occlusions, and simultaneously produces consistent ground plane and contact outputs.

Limitations & Future Work. HuMoR leaves ample room for future studies. The static camera and ground plane assumptions are reasonable for indoor scenes but true *in-the-wild* operation demands methods handling dynamic cameras and complex terrain. Our rather simplistic contact model should be upgraded to capture scene-person interactions for improved motion and scene perception. Lastly, we plan to *learn* motion estimation directly from partial observations which will be faster than TestOpt and enable *sampling* multiple plausible motions rather than relying on a single local minimum.

Acknowledgments. This work was supported by the Toyota Research Institute ("TRI") under the University 2.0 program, grants from the Samsung GRO program and the Ford-Stanford Alliance, a Vannevar Bush faculty fellowship, NSF grant IIS-1763268, and NSF grant CNS-2038897. TRI provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity.

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7144–7153, 2019. 3

[2] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5223–5232, 2020. 6

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 2, 4, 6

[4] Andreas Baak, Meinard Müller, Gaurav Bharaj, Hans-Peter Seidel, and Christian Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *Consumer Depth Cameras for Computer Vision*, pages 71–98. Springer, 2013. 2

[5] Albert E Beaton and John W Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974. 5

[6] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1171–1179, Cambridge, MA, USA, 2015. MIT Press. 4

[7] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020. 2

[8] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016. 2, 4, 5

[9] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016*, pages 10–21. Association for Computational Linguistics (ACL), 2016. 4

[10] Matthew Brand and Aaron Hertzmann. Style machines. In *ACM SIGGRAPH*, pages 183–192, July 2000. 2

[11] Marcus A. Brubaker, David J. Fleet, and Aaron Hertzmann. Physics-based person tracking using the anthropomorphic walker. *IJCV*, (1), 2010. 1, 2

[12] Marcus A. Brubaker, Leonid Sigal, and David J. Fleet. Estimating contact dynamics. In *Proc. ICCV*, 2009. 1, 2

[13] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 2, 6

[14] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6

[15] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8648–8657, 2019. 1

[16] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, pages 20–40. Springer, 2020. 2

[17] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 3

[18] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. *Advances in Neural Information Processing Systems (NeurIPS)*, 32:12949–12961, 2019. 2

[19] Ahmed Elgammal and Chan-Su Lee. Separating style and content on a nonlinear manifold. In *IEEE Conf. Comp. Vis. and Pattern Recognition*, pages 478–485, 2004. Vol. 1. 2

[20] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. Real time motion capture using a single time-of-flight camera. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 755–762. IEEE, 2010. 2

[21] S. Geman and D. McClure. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, 52(4):5–21, 1987. 5

[22] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *European Conference on Computer Vision*, pages 768–784. Springer, 2020. 2

[23] Saeed Ghorbani, Calden Wloka, Ali Etemad, Marcus A Brubaker, and Nikolaus F Troje. Probabilistic character motion synthesis using a hierarchical deep latent variable model. In *Computer Graphics Forum*, volume 39. Wiley Online Library, 2020. 2

[24] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 2

[25] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 2

[26] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 1

[27] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. In *28th British Machine Vision Conference*, 2017. 2

[28] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, pages 2282–2292, 2019. 2, 5, 6, 7, 8

[29] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718, 2021. 2

[30] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 2

[31] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 2

[32] Nicholas R. Howe, Michael E. Leventon, and William T. Freeman. Bayesian reconstruction of 3D human motion from single-camera video. In *Advances in Neural Information Processing Systems 12*, pages 820–826, 2000. 2

[33] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *2017 International Conference on 3D Vision*, pages 421–430. IEEE, 2017. 2, 4, 6

[34] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition*, 2018. 2

[35] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2, 6

[36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 4

[37] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 1, 2, 6, 8

[38] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. *arXiv preprint arXiv:2104.08527*, 2021. 2

[39] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings International Conference on Computer Vision (ICCV)*, pages 2252–2261. IEEE, Oct. 2019. ISSN: 2380-7504. 2

[40] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. In *ACM Transactions on Graphics 21(3), Proc. SIGGRAPH*, pages 473–482, July 2002. 2

[41] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017. 2

[42] Yan Li, Tianshu Wang, and Heung-Yeung Shum. Motion texture: A two-level statistical model for character motion synthesis. In *ACM Transactions on Graphics 21(3), Proc. SIGGRAPH*, pages 465–472, July 2002. 2

[43] Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. Estimating 3d motion and forces of person-object interactions from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[44] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. Character controllers using motion vaes. In *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, volume 39. ACM, 2020. 2, 3, 4, 6, 7, 8

[45] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4450–4459, 2019. 6

[46] C. Karen Liu, Aaron Hertzmann, and Zoran Popović. Learning physics-based motion style with nonlinear inverse optimization. *ACM Trans. Graph*, 2005. 2

[47] Miao Liu, Dexin Yang, Yan Zhang, Zhaopeng Cui, James M Rehg, and Siyu Tang. 4d human body capture from egocentric video via 3d scene grounding. *arXiv preprint arXiv:2011.13341*, 2020. 2

[48] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2, 3

[49] James Lucas, George Tucker, Roger B Grosse, and Mohammad Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 4

[50] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[51] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, 2019. 2, 6

[52] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 39(4):82–1, 2020. 1, 2

[53] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 1, 2

[54] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J. Mitra. iMapper: Interaction-guided scene mapping from monocular videos. *ACM SIGGRAPH*, 2019. 2, 6, 7, 8

[55] Dirk Ormoneit, Hedvig Sidenbladh, Michael J. Black, and Trevor Hastie. Learning and tracking cyclic human motion. In *Advances in Neural Information Processing Systems 13*, pages 894–900, 2001. 1, 2

[56] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems*, 2017. 5

[57] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1, 2, 5, 6

[58] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018. 2

[59] Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. Modeling human motion with quaternion-based neural networks. *International Journal of Computer Vision*, pages 1–18, 2019. 3

[60] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 2

[61] Vladimir Pavlović, James M. Rehg, and John MacCormick. Learning switching linear models of human motion. In *Advances in Neural Information Processing Systems 13*, pages 981–987, 2001. 1, 2

[62] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Trans. Graph*, 2018. 1, 2

[63] Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 6

[64] Chris Rockwell and David F Fouhey. Full-body awareness from partial observations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 522–539, 2020. 2

[65] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017. 3

[66] Charles Rose, Michael F. Cohen, and Bobby Bodenheimer. Verbs and adverbs: Multidimensional motion interpolation. *IEEE Computer Graphics and Applications*, 18(5):32–40, 1998. 2

[67] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[68] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Transactions on Graphics (TOG)*, 40(1):1–15, 2020. 2

[69] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Trans. Graph.*, 39(6), Nov. 2020. 1, 2

[70] Hedvig Sidenbladh, Michael J. Black, and David J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *ECCV*, pages 702–718, 2000. Part II. 1, 2

[71] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010. 6

[72] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 3483–3491. Curran Associates, Inc., 2015. 2, 3, 4

[73] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209–1, 2019. 2

[74] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5349–5358, 2019. 2

[75] Graham W. Taylor, Geoffrey E. Hinton, and Sam T. Roweis. Modeling human motion using binary latent variables. In *Proc. NIPS*, 2007. 1, 2

[76] Raquel Urtasun, David J. Fleet, and Pascal Fua. 3D people tracking with Gaussian process dynamical models. In *IEEE Conf. Comp. Vis. & Pattern Rec.*, pages 238–245, 2006. Vol. 1. 1, 2

[77] Raquel Urtasun, David J. Fleet, and Pascal Fua. Temporal motion models for monocular and multiview 3D human body tracking. *CVIU*, 104(2):157–177, 2006. 1, 2

[78] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 2

[79] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129(7):2264–2287, 2021. 1

[80] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pages 109–117, 2017. 1

[81] Jack M. Wang. Gaussian process dynamical models for human motion. Master's thesis, University of Toronto, 2005. 2

[82] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 4

[83] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10974, 2019. 2

[84] Dongseok Yang, Doyeon Kim, and Sung-Hee Lee. LoBSTr: Real-time Lower-body Pose Prediction from Sparse Upper-body Tracking Signals. *Computer Graphics Forum*, 2021. 2

[85] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020. 6

[86] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[87] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision*, pages 465–481. Springer, 2020. 2

[88] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018. 2, 6

[89] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7114–7123, 2019. 2

[90] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[91] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021. 3