Full-Body Motion from a Single Head-Mounted Device: Generating SMPL Poses from Partial Observations

Andrea Dittadi[1],* Sebastian Dziadzio[2], Darren Cosker[3], Ben Lundell[2], Tom Cashman[2], Jamie Shotton[2]
[1]Technical University of Denmark [2]Microsoft [3]University of Bath

Abstract

The increased availability and maturity of head-mounted and wearable devices opens up opportunities for remote communication and collaboration. However, the signal streams provided by these devices (e.g., head pose, hand pose, and gaze direction) do not represent a whole person. One of the main open problems is therefore how to leverage these signals to build faithful representations of the user. In this paper, we propose a method based on variational autoencoders to generate articulated poses of a human skeleton based on noisy streams of head and hand pose. Our approach relies on a model of pose likelihood that is novel and theoretically well-grounded. We demonstrate on publicly available datasets that our method is effective even from very impoverished signals and investigate how pose prediction can be made more accurate and realistic.

1. Introduction

Head-mounted and wearable devices are steadily increasing in availability and maturity. These technologies open up opportunities to build tools for remote communication and collaboration that are human-centred, and which allow us to work in the way we naturally interact when we meet in person [27, 50, 62]. Mixed reality devices, such as Microsoft HoloLens, allow 3D content to be displayed and viewed in physical space with local or remote collaborators, all sharing a single coordinate system and spatial context. However, to communicate effectively with remote collaborators, there is a significant and largely unsolved challenge to build faithful representations of the motion of a person wearing such a headset from only head-worn sensors. There is a high perceptual bar to meet if we are to trust such a system to represent ourselves, as we are attuned to motion that does not look human [44].

To sense the motion and actions of a user, devices such as HoloLens provide a variety of signal streams derived using computer vision; these include the location and orientation of the head-mounted device (HMD) relative to a world coordinate system, hand pose (location and orientation of the user's hands relative to the HMD) and even eye tracking signals [66]. These signal streams provide invaluable information but they do not represent a whole person. Furthermore, while each individual stream has its own failure rate due to detection or tracking errors, the combination of all the streams has a much higher *compound* failure rate, as a failure in any one subsystem can result in non-human behaviour that breaks the trust and understanding required for effective communication.

While the possibility of estimating the full body pose of a person using egocentric views is an attractive prospect on future devices [64, 52], no currently available consumer device has suitable embedded cameras. On a wearable device, each additional camera is costly in terms of power and thermal dissipation [35], and so it is advantageous for motion prediction systems to require as few cameras as possible. Even if future devices provide egocentric body tracking cameras, there will be an ongoing need to allow full-body representations for users of legacy or lower-power devices. Solutions that predict body motion from external cameras mounted on an interacting person are also promising [46], but are limited to cases were there are multiple people interacting and participants are always visible.

To address this challenge, we require a model of human motion that is conditioned on limited low-level inputs and provides plausible inferred body poses while staying responsive to the signal streams. In this paper, we address an important sub-problem: reconstructing the articulated pose of a human skeleton from noisy streams of head and hand pose. We use the variational autoencoder (VAE) framework [30, 51], which allows us to decompose the problem into a *generative* model of human pose, with an *inference* model that maps input signals into the learned latent embedding.

Our primary contribution is to show how to make this framework effective even from *very impoverished signals*: in our case the three orthogonal coordinate frames provided by a head and hand tracker. Our secondary contribution is to formulate a model of pose likelihood which is factorized

*Work done during an internship at Microsoft Research. Correspondence to: adit@dtu.dk into approximately Gaussian models for each of the joints in our skeletal model; this leads to an objective function that is novel and theoretically well-grounded by the manifold of poses for each joint in the special Euclidean group SE(3). Finally, we show that the accuracy of pose prediction can be improved in two main ways: first, by using a generative model that is pretrained on full body poses, and second, by providing a temporal history of head and hand poses to the inference model.

2. Related Work

We present a deep generative model that is able to produce diverse and natural sequences of human body poses within the constraints imposed by the head and hand tracking signals. Our work therefore lies at the intersection of motion control, motion prediction, and egocentric pose estimation. However, unlike motion control approaches, we do not have a clean future trajectory of the root joint or floor path. Unlike motion prediction, our problem is focused on predicting the present pose from incomplete data, rather than the far future from complete previous knowledge. And unlike the egocentric pose estimation literature, we assume that the only available signals are the three coordinate frames that give the position and orientation of the head and hands.

Motion control. Motion control is the problem of generating plausible and varied motion given control input: a temporally dense, usually user-defined signal, such as direction or trajectory. Traditional approaches use graph-based search algorithms to find suitable segments in the motion database and concatenate them to produce the desired sequence [5, 55]. While effective, flexible—and in the case of nearest-neighbour motion matching [11]—widely used, search-based techniques are constrained by their memory requirements, which scale linearly with the amount of data. Machine learning techniques address this problem by distilling the motion database into a statistical model with bounded runtime memory and computation requirements. Recent approaches rely on phase-functioned or motion matching networks to create full body animations from e.g. game controller input [23, 22]. However, the requirement to provide reliable future direction and position vectors based on the pelvis make motion matching unsuitable for HMDs.

Motion prediction. Work in the area of motion and pose prediction attempts to forecast future body poses from observed past data. Deterministic approaches treat the problem as a regression task with a single correct solution. These include feed-forward networks [10, 41], convolutional autoencoders [24, 17], recurrent neural networks [42, 15, 16, 25], often with adversarial components [17, 33], or reinforcement learning [67]. Other approaches include probabilistic models, such as conditional restricted Boltzmann machines [61, 60], variational autoencoders [2, 3, 69, 73], and normalizing flows [20]. In general, research in this area assumes that body poses observed in the past are complete—with no missing joints. In addition, it ignores the notion of a motion controller. Both of these conditions make our problem distinct, i.e. body pose prediction should be from very sparse signals (i.e. the head and hands only), and these signals provide motion cues to guide prediction.

Egocentric pose estimation. Capturing full 3D body motion from head-mounted cameras presents a significant challenge, mostly due to self-occlusions. Most approaches for egocentric motion estimation are based on head-mounted cameras and reconstruct upper body motion (hands, arms and torso) [14, 38, 12, 70, 53]. The approach proposed by Jiang and Grauman [26] reconstructs full-body pose by estimating egomotion from the scene observed from a camera placed on the user's chest. Yuan and Kitani [71, 72] propose a method based on a control-based representation of humanoid motion. The setup in [52] consists of a pair of fisheye cameras mounted on a helmet, which capture the whole body. More recent methods for full-body pose estimation from more compact head-mounted devices have proven successful [68, 64, 63]. However, despite severe self-occlusions, the information available in these cases is still richer than the head and hands signal we consider in this paper.

Other related work. In the context of reconstructing poses from partial data, a related problem is to recover 3D poses from 2D images [4, 9, 32, 43, 47, 54, 59, 65, 74], with some methods explicitly targeting ambiguous scenarios [1, 8, 34, 56, 57, 58]. In particular, [8] considers the setting where the images present heavy occlusions, which is in some sense more closely related to our setting. Other related but distinct settings are motion in-betweening and infilling [18, 19, 28] where the task in broad terms is to fill temporal gaps in observed sequences. Regarding learning a VAE on human poses, [49] proposes a similar approach, but has a different context and goal (learning a human pose prior), additional terms in the loss function that do not derive from the VAE formulation [49, Sec. 3.3], and a different likelihood function (cf. Eq. (4) and Appendix A in this paper).

3. Background and Notation

The SMPL model of human shape and pose. We use the SMPL model [37] to represent articulated human body pose. In particular, we represent the pose as a configuration of 22 joints arranged in a kinematic tree that defines the coarse structure of the human skeleton;[1] see Figure 1. Each

[1] While the original SMPL model includes 24 joints, we exclude the two joints that model hand bending, leaving a full articulated model of body

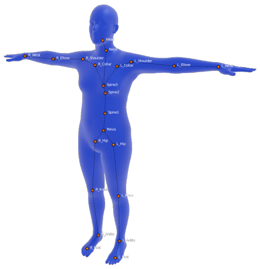

Figure 1. The SMPL model with the 22 bones used in this work, left: in the neutral pose, and right: when posed by rotations Ω.

of these 22 joints has a six degrees-of-freedom (6-DoF) pose given by an element of the Lie group $\mathrm{SE}(3)$, so we think of a full body pose as an element of $\mathrm{SE}(3)^{22}$. The kinematic tree defines the position of any joint j in a full body pose by the shape of the skeleton in the neutral pose, and the rotation $\omega_j \in \mathrm{SO}(3)$ of the joint *relative* to its parent in the tree. In this way, we are able to easily move between local and global representations of full body pose, with local pose $\Omega = \{\omega_j\}_{j=1}^{22}$ and global pose $\mathbf{x} = G(\Omega)$ calculated by recursively evaluating the pose of each joint in the kinematic tree. We are focused entirely on learning the relative poses Ω of the joints and ignore the question of shape in this work. We use the mean shape of the SMPL model in all our experiments, and leave the problem of skeletal shape inference for future work.

Variational autoencoders. In a latent variable model, we suppose that an observable random variable $\mathbf{x}$ can be obtained as a transformation of an unobserved (or latent) random variable $\mathbf{z}$. Writing $p_\theta(\mathbf{x})$ and $p_\theta(\mathbf{z})$ for the distributions of $\mathbf{x}$ and $\mathbf{z}$, respectively, with parameters θ, we have the latent variable model: $p_\theta(\mathbf{x}) = \int_\mathbf{z} p_\theta(\mathbf{z}) p_\theta(\mathbf{x} \,|\, \mathbf{z}) d\mathbf{z}$. In most practical cases, optimizing the marginal likelihood $p_\theta(\mathbf{x})$ is not directly possible due to the intractability of the integral. Thus, in variational inference the quantity being maximized is typically the Evidence Lower Bound (ELBO):

$$\begin{aligned} \log p_\theta(\mathbf{x}) &= \log \int_\mathbf{z} p_\theta(\mathbf{x} \,|\, \mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z} \\ &= \log \mathbb{E}_{q_\phi(\mathbf{z})} \left[\frac{p_\theta(\mathbf{x} \,|\, \mathbf{z}) p_\theta(\mathbf{z})}{q_\phi(\mathbf{z})} \right] \\ &\geq \mathbb{E}_{q_\phi(\mathbf{z})} \left[\log \frac{p_\theta(\mathbf{x} \,|\, \mathbf{z}) p_\theta(\mathbf{z})}{q_\phi(\mathbf{z})} \right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z})} \left[\log p_\theta(\mathbf{x} \,|\, \mathbf{z}) \right] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}) \| p_\theta(\mathbf{z})) \end{aligned} \tag{1}$$

which bounds the log likelihood from below for *any* distribution $q_\phi(\mathbf{z})$ parameterized by ϕ. In the last line, after changing the signs to obtain a loss function, the first term can be interpreted as expected reconstruction loss, and the second as a regularizer for the latent representations $\mathbf{z}$.

Notation. We denote a full body pose by $\mathbf{x} \in \mathrm{SE}(3)^{22}$, partitioned as follows:

$$\mathbf{x} = [\mathbf{x}_{\mathrm{b}}, \mathbf{x}_{\mathrm{hh}}] \tag{2}$$

where $\mathbf{x}_{\mathrm{hh}} \in \mathrm{SE}(3)^3$ denotes head and hands, and $\mathbf{x}_{\mathrm{b}} \in \mathrm{SE}(3)^{19}$ the rest of the body. When considering the temporal dimension, we add the time step index as a superscript. For example, $\mathbf{x}_{\mathrm{hh}}^t$ is the head and hands observation at time t, and $\mathbf{x}_{\mathrm{hh}}^{t_1:t_2}$ is a sequence of head and hands observations between t_1 and t_2, both included, with $t_1 \leq t_2$.

4. Methods

A generative model of human poses. We model human poses $\mathbf{x}^t$ with a latent variable model:

$$p_\theta(\mathbf{x}^t) = \int_{\mathbf{z}^t} p_\theta(\mathbf{x}^t \,|\, \mathbf{z}^t) p(\mathbf{z}^t) d\mathbf{z}^t \tag{3}$$

with θ being a parameter vector.[2] We assume that a pose $\mathbf{x}$ arises from a generative process where a latent variable $\mathbf{z}$ is sampled from a fixed prior distribution $p(\mathbf{z})$, and then a pose is sampled from a conditional distribution $p_\theta(\mathbf{x} \,|\, \mathbf{z})$. We choose the prior to be an isotropic Gaussian with unit variance: $p(\mathbf{z}) = \mathcal{N}(\mathbf{z};\, \mathbf{0}, \mathbf{I})$, and parameterize the likelihood $p_\theta(\mathbf{x} \,|\, \mathbf{z})$ with a decoder network $d_\theta(\mathbf{z})$ as follows.

The random variable $\mathbf{x} \in \mathrm{SE}(3)^{22}$ represents the 6-DoF pose of all 22 joints in the SMPL model [37] relative to some choice of global coordinate frame. As is standard [31], we assume conditional independence of our generative model $p_\theta(\mathbf{x} \,|\, \mathbf{z})$. Thus, we have a factorization

$$p_\theta(\mathbf{x} \,|\, \mathbf{z}) = \prod_{j=1}^{22} p_\theta(P_j \,|\, \mathbf{z})$$

where P_j is the 6-DoF pose of joint j. Note that $p_\theta(P_j \,|\, \mathbf{z})$ is a probability distribution over the Lie group $\mathrm{SE}(3)$. Let $(\boldsymbol{\mu}_{j,\theta}(\mathbf{z}), \boldsymbol{\sigma}^2_{j,\theta}(\mathbf{z}))$ represent the component of $d_\theta(\mathbf{z})$ corresponding to the jth joint. Then we take

$$p_\theta(P_j \,|\, \mathbf{z}) \propto \exp\left\{ -\frac{1}{2} d_{\mathrm{SE}(3)} \left(P_j, \boldsymbol{\mu}_{j,\theta}(\mathbf{z}); \boldsymbol{\sigma}^2_{j,\theta}(\mathbf{z}) \right)^2 \right\} \tag{4}$$

where $d_{\mathrm{SE}(3)}(A, B; \Sigma)$ represents the left-invariant geodesic distance between $A, B \in \mathrm{SE}(3)$ arising from the quadratic form Σ defined on $\mathfrak{se}(3) \simeq \mathbb{R}^6$. In this work, we take $\boldsymbol{\sigma}^2_{j,\theta}$ to be constant and independent of both j and θ for simplicity. Specifically, we choose

$$\boldsymbol{\sigma}^2_{j,\theta} = \begin{bmatrix} s_1^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & s_2^2 \mathbf{I} \end{bmatrix}, \tag{5}$$

pose without hand articulation

[2] When no confusion arises, we will drop the dependence on time t.

where $s_1 = 2$ centimeters, $s_2 = 0.1$ radians, and $\mathbf{I}$ is the 3×3 identity matrix. See Appendix A for an explanation of how to interpret $p_\theta(P_j \mid \mathbf{z})$ as the product of two independent (nearly) Gaussian distributions.

We note that there does not exist a bi-invariant geodesic distance on $\mathrm{SE}(3)$ [48, Theorem 1], so one must make a choice of a left-invariant metric or a right-invariant metric. The former corresponds to a metric which is invariant to global coordinate changes, while the latter corresponds to a metric which is invariant to local coordinate changes. In our setting, we choose the left-invariant metric because our choice of local coordinates is fixed by the SMPL model, while our choice of global coordinates is arbitrary.

An inference model of latent space. Following the variational autoencoder framework [30, 51], we define an inference model with parameters ϕ that parameterizes the approximate posterior of the latent variables, $q_\phi(\mathbf{z} \mid f(\mathbf{x}))$. Here $f(\mathbf{x})$ can be any information that might help inference, and it is reasonable to assume that it should depend on $\mathbf{x}$. Since we are interested in generating full body poses from partial observations (in particular, from head and hands only), we train special inference models that infer plausible values of $\mathbf{z}$—*i.e.*, that will generate plausible poses through the generative model—from those partial observations. Although the classic VAE setting corresponds to $f(\mathbf{x}) = \mathbf{x}$, this is not necessary (see Eq. (1)). We define the approximate posterior as a Gaussian with diagonal covariance:

$$q_\phi(\mathbf{z} \mid f(\mathbf{x})) = \mathcal{N}\left(\mathbf{z};\, \boldsymbol{\mu}_\phi(f(\mathbf{x})), \operatorname{diag}\left(\boldsymbol{\sigma}^2_\phi(f(\mathbf{x}))\right)\right) \quad (6)$$

where the mean and scale parameters, $\boldsymbol{\mu}_\phi(f(\mathbf{x}))$ and $\boldsymbol{\sigma}^2_\phi(f(\mathbf{x}))$, are given by an encoder network $e_\phi(f(\mathbf{x}))$. We describe the choices of f we use in this paper below.

Training encoder and decoder networks. The inference and generative models (*i.e.*, the encoder and decoder networks) are typically trained end-to-end by maximizing a Monte Carlo approximation of the ELBO (1) via stochastic gradient ascent. Since the prior and approximate posterior are multivariate Gaussians, the KL divergence term can be computed analytically. On the other hand, the reconstruction term has to be estimated by sampling from the inference model. Furthermore, in order to control the trade-off between the two terms of the loss, we follow the β-VAE framework [21] and scale the KL divergence by a (fixed) non-negative scalar β. The objective function then becomes:

$$\mathcal{L}(\theta, \phi) = \frac{1}{S}\left(\sum_{s=1}^{S} \log p_\theta\big(\mathbf{x} \mid \mathbf{z}^{(s)}\big)\right) - \beta D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid f(\mathbf{x})) \,\|\, p(\mathbf{z})) \quad (7)$$

with $\mathbf{z}^{(s)} \sim q_\phi(\mathbf{z} \mid f(\mathbf{x}))$ for $s = 1, \ldots, S$. Using Equations (4) and (5) and the results from Appendix A, we have that $\log p_\theta\big(\mathbf{x} \mid \mathbf{z}\big)$ is a weighted sum of the squared positional and angular errors at each joint.

Inference from incomplete data. As we are interested in full body pose prediction from head and hands data, we consider the case where inference input depends only on $\mathbf{x}_{\mathrm{hh}}$. The first obvious case is $f_1(\mathbf{x}) = \mathbf{x}^t_{\mathrm{hh}}$ where the input is simply the head and hands observation. Then we can note that optimizing the generative model requires Monte Carlo samples from the inference model $q_\phi(\mathbf{z} \mid f(\mathbf{x}))$. If the inference model has less information available, the quality of inference—how close the inference model is to the true posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$—might be impaired, which could, in turn, negatively affect the training of the generative model. One way to address this issue is to pre-train a regular VAE using full body poses as inputs to the inference model, freeze the generative model, and define a new inference model conditioned on incomplete information. We can then optimize the ELBO using the pre-trained generative model. Since head and hands observations from previous time steps might help inference by resolving ambiguities arising from the missing information we also explore the effect of including this information in $f(\mathbf{x})$. In summary, we consider five settings:

- Model 0: $f_0(\mathbf{x}) = \mathbf{x}^t$, *i.e.*, the classic VAE setting, which is inapplicable in practical contexts given the requirement for full body pose as input.
- Model 1: $f_1(\mathbf{x}) = \mathbf{x}^t_{\mathrm{hh}}$.
- Model 2: $f_2(\mathbf{x}) = \mathbf{x}^{t-T:t}_{\mathrm{hh}}$, which gives the inference model a history of head and hand observations.
- Model 3: $f_3(\mathbf{x}) = \mathbf{x}^t_{\mathrm{hh}}$, as in model 1, but with the generative model $d_\theta(\mathbf{z})$ pre-trained as a standard VAE (model 0).
- Model 4: $f_4(\mathbf{x}) = \mathbf{x}^{t-T:t}_{\mathrm{hh}}$, as in model 2, but with the generative model $d_\theta(\mathbf{z})$ pre-trained as a standard VAE (model 0).

5. Experiments

This section describes how we test models trained in these five settings for the target problem: generating plausible full-body pose of users wearing an HMD.

5.1. Data

Datasets. For training data we need sequences of full body motion that capture the target poses and temporal behaviour we want to learn. For this purpose, it's convenient to use motion capture data that has been fitted to express the pose in each frame with the SMPL model. Such a dataset is provided by Mahmood *et al.* in the form of AMASS [39]. We use the `KIT` [40], `MPI_HDM05` [45] and `CMU` [13] datasets from

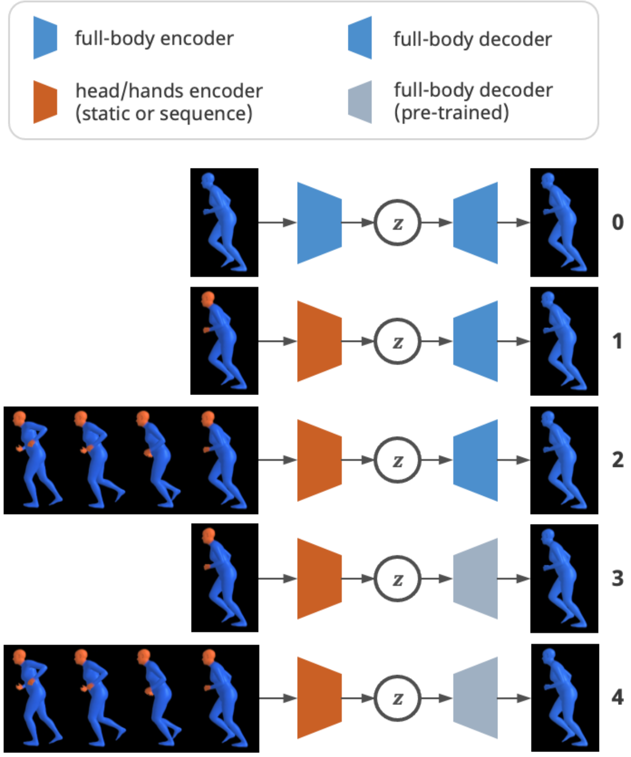
full-body encoder
full-body decoder
head/hands encoder (static or sequence)
full-body decoder (pre-trained)
z
0
1
2
3
4

Figure 2. **Overview of the five settings considered in this paper** (numbers on the right). Model 0 is the standard VAE, and $\mathbf{z}$ is a representation of the reconstructed pose on the right. All other models have only head and hand poses available for inference of $\mathbf{z}$ (highlighted in orange in the meshes on the left). Models 1 and 3 take static head and hands observations as input, while models 2 and 4 have access to previous frames. Models 3 and 4 do not train a decoder from scratch, but use a pre-trained decoder from model 0.

AMASS, and process each by downsampling each sequence by an integer factor, such that the final framerate is as close as possible to 30 Hz, the minimum framerate across these datasets. Since we also want to use sequences as input, we fix a sequence length $L = 16$ for training, corresponding to approximately 0.5 seconds, and $L_{\text{test}} = 64$ for testing and visualization, corresponding to approximately 2 seconds. During preprocessing we discard all sequences shorter than L_{test}. Then we randomly choose approximately 5% of the remaining sequences from each of the three datasets, and hold them out for testing. From each training and test sequence we then extract subsequences of length L and L_{test}, respectively, with a sliding window that shifts by 4 frames at a time. The resulting training and test set comprise 278,431 sequences of length $L = 16$, and 11,797 sequences of length $L_{\text{test}} = 64$, respectively.

Data representations. Each frame is stored as the local rotations Ω of the 22 joints in the SMPL body model, in the axis-angle representation. For simplicity, we set the origin of the coordinate system to the position of the pelvis, which is the root of the kinematic tree. When loading the datasets, we convert the axis-angle vectors to rotation matrices, and rotate the pelvis of each frame around the vertical axis such that, in the final frame of each sequence, the projection onto the ground plane of the direction faced by the pelvis is always the same. In this way the generative model is trained to only produce poses facing a specific direction, thus enforcing rotation invariance around the vertical axis. This makes sense in the context of deployment to available HMDs, which commonly provide an inertial measurement unit that can output an 'up' (gravity) vector, about which we can rotate the input data to perform a similar normalization.

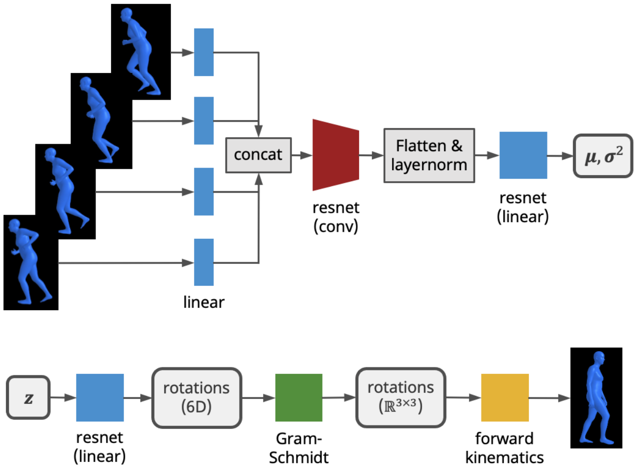

Figure 3. **Simplified diagram of encoder from sequences (top) and decoder (bottom).** We omit the encoder from static poses as it simply consists of a residual network. In the encoder, input body poses are represented as global translations and rotations of all joints (in the fully observable case, *i.e.*, in model 0) or only head and hands (in all other cases). Each resnet consists of a number of residual blocks, each containing 2 linear or convolutional layers. The total number of residual blocks is the same in the pose encoder, sequence encoder, and decoder.

Note that by using the *left*-invariant geodesic on $\mathrm{SE}(3)$ as our log-likelihood, the loss computed during training is independent of these data normalization choices.

5.2. Model Architecture

Here we describe the encoder and decoder architectures used in our experiments. For further implementation details we refer to Appendix B.

Decoder. See Fig. 3 (bottom) for a simplified diagram of the decoder. The first component of the generative model is a neural network that takes as input a latent vector $\mathbf{z}$ of size d and outputs a vector of size $22 \cdot 6 = 132$ encoding a continuous 6D representation [75] of the relative rotation of all 22 joints. This neural network consists of a linear layer that maps $\mathbf{z}$ onto a 256-dimensional space, followed by a number of residual blocks and a final linear layer that outputs a 132-dimensional vector. Following Zhou *et al.* [75], this vector is then transformed into 22 rotation matrices $\hat{\Omega} \in \mathrm{SO}(3) \subset \mathbb{R}^{3\times 3}$ via Gram-Schmidt orthogonalization and finally, using the forward kinematics $G(\hat{\Omega})$ of the SMPL model, into 22 matrices that represent $\mathbf{x}$: the global translations and rotations of the joints.

Encoder from static poses. The inference model takes as input the $\mathrm{SE}(3)$ global translations and rotations of each joint, each represented by a vector in $\mathbb{R}^3$ and a 3×3 matrix, which are both flattened to give a vector of size 12. The number of input joints is either 3 (in the head and hands case) or 22, thus the input has size $3 \cdot 12 = 36$ or $22 \cdot 12 = 264$ respectively. Similarly to the decoder described above, this input is linearly transformed into a 256-dimensional vector, passed through a series of residual blocks, and finally transformed with a linear layer into a vector of size $2d$ representing the mean and log variance of the approximate posterior.

Encoder from sequences. See Fig. 3 (top) for a simplified diagram of this encoder. The inputs to the encoder in this case are the $\mathrm{SE}(3)$ global translations and rotations of each joint at each time step of the sequence, represented as in the static case above. Note that in the sequential case we only consider head and hands input, thus the input has size $36 \times L$, with L the sequence length. In this encoder we use 1D convolutional layers where along the channel dimension we concatenate a vector with a linear positional encoding in $[-1, 1]$, similar to the approach described by Liu *et al.* [36].

The first component of the encoder is a linear mapping that maps each input frame independently to a vector of size 128. This is followed by a series of residual blocks similarly to the decoder, with the difference that most of the residual blocks are based on 1D convolutional layers rather than linear layers. The output is again a vector of size $2d$ that parameterizes the variational distribution.

5.3. Results

Hyperparameter sweep. We train the 5 classes of models described above, and for each class we vary the following hyperparameters:

- β in the objective function (7) takes values in $\{0.0001, 0.01, 0.1, 1, 5\}$. Note that when $\beta \ll 1$ the VAE roughly approximates a deterministic autoencoder.
- The latent space dimensionality is $d \in \{15, 30, 60\}$
- Encoder and decoder always have the same number of residual blocks, which is either 3 or 6.

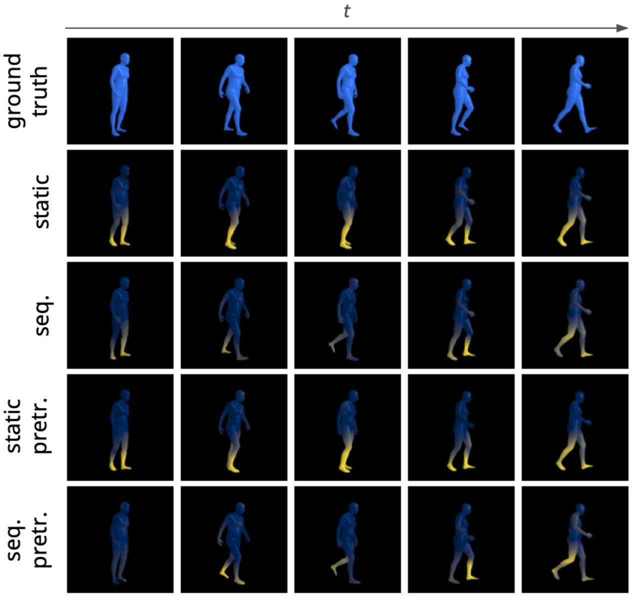

Figure 4. **Ground truth and reconstructions of a walking sequence by different models.** First row: ground truth. In the other rows, models 1 to 4 are reported in order. The meshes are color-coded according to position error.

Overall we have 30 different hyperparameter combinations and 5 model types, thus 150 models in total. From model 0 we only need the pre-trained decoder, and in the following evaluation we consider models 1 to 4, totalling 120 models.

Qualitative results. The datasets used in this work encompass a wide variety of motions, from standing or walking to dancing or playing sports. Given our focus on typical use cases for HMDs, we would like to get a sense of how our approach performs when a person is engaged in common everyday behaviours. Fig. 4 shows a ground truth walking sequence along with the output sequences reconstructed by the proposed approach (models 1 to 4) given only head and hands observations. We qualitatively observe (see also Supplementary Material) that including past information seems to lead to more accurate and natural generated sequences. Furthermore, perhaps unsurprisingly, the main contribution to the average position error appears to come from the leg joints (knees, ankles, and feet).

Quantitative results. We now quantitatively evaluate the trained models on a broader test set consisting of randomly selected sequences from the datasets mentioned earlier. We compute the following metrics:

- **Average position error**: The average Euclidean distance between the ground truth and predicted 3D joint positions. For a sequence of length T and a set of joints $\mathcal{J}$, this metric is defined as:

$$\frac{1}{T \cdot |\mathcal{J}|} \sum_{t=1}^{T} \sum_{j \in \mathcal{J}} \|\hat{\mathbf{x}}_j^t - \mathbf{x}_j^t\|_2 \tag{8}$$

  where $\mathbf{x}_j^t$ and $\hat{\mathbf{x}}_j^t$ denote the true and predicted (3-dimensional) global position of joint j at time step t, respectively.

- **Average velocity error**: The average Euclidean distance between the ground truth and predicted 3D joint velocities. This metric is defined as:

$$\frac{1}{(T-1) \cdot |\mathcal{J}|} \sum_{t=2}^{T} \sum_{j \in \mathcal{J}} \|\hat{\mathbf{v}}_j^t - \mathbf{v}_j^t\|_2 \tag{9}$$

  where $\mathbf{v}$ denotes joint velocity:

$$\mathbf{v}_j^t = \frac{\mathbf{x}_j^t - \mathbf{x}_j^{t-1}}{\Delta t} \tag{10}$$

  (and similarly for $\hat{\mathbf{v}}$) and the sampling period Δt is $1/30$ seconds.

- **Average acceleration**: The average magnitude of the predicted joint accelerations. This metric measures the smoothness of the generated pose sequence [72], and is defined as

$$\frac{1}{(T-2) \cdot |\mathcal{J}|} \sum_{t=3}^{T} \sum_{j \in \mathcal{J}} \|\hat{\mathbf{a}}_j^t\|_1 \tag{11}$$

  where $\hat{\mathbf{a}}$ denotes joint acceleration

$$\hat{\mathbf{a}}_j^t = \frac{\hat{\mathbf{v}}_j^t - \hat{\mathbf{v}}_j^{t-1}}{\Delta t}\,. \tag{12}$$

Here $\mathcal{J}$ denotes a set of joints over which the above metrics are averaged. We report results for the full body as well as for the legs only. Note that predicting the motion of the lower body represents the greatest challenge for all methods on this problem, as our input signals appear only from the upper body.

Fig. 5 shows position and velocity errors averaged over the full body. The label ‘static’ denotes inference models that take as input head and hands information at the current time step only, while ‘sequence’ indicates that the input includes head and hands history. Models marked as ‘pretrained’ indicate that the decoder paired with the inference model has been pretrained with an encoder of full-body pose (model 0).

Although these metrics do not seem to be particularly discriminative in this case, this improves when considering the same metrics on leg joints only, as shown in Fig. 6. Here the benefit of a deeper architecture is more evident, and including past head and hands information in the input appears to consistently improve both position and velocity errors. While the best option in terms of position error is to train the decoder from scratch along with a sequential inference model (model 2), pretraining the decoder seems to lead to the lowest average leg velocity error.

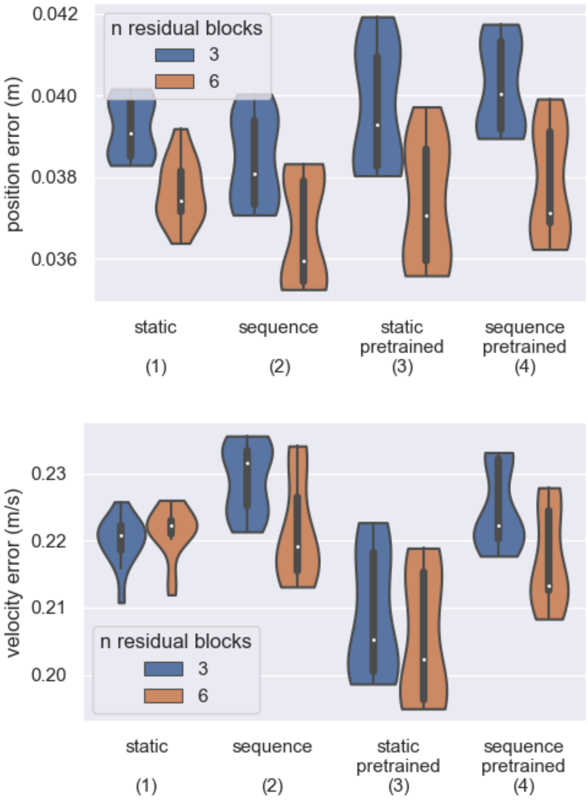

Figure 5. Average joint **position** (top) and **velocity** (bottom) errors on test sequences. Average over **full body**.

In Fig. 7 we look at average joint accelerations of the reconstructed sequences, both on the full body and on legs only. This can be considered a proxy for smoothness, which is a relevant property of the generated sequences, considering our main use cases. These plots highlight a trend towards lower average joint accelerations for pretrained decoders and sequential input. In particular, training a VAE from scratch with an inference model that takes as input head and hands at the current time step appears to lead to high acceleration, and therefore less smooth sequences. This trend is particularly clear when averaging over legs only.

Finally, note that increasing the depth of the networks seems to improve results in terms of position error, but not necessarily velocity error. This is not particularly surprising, as deeper networks tend to work better and we are optimizing position error, not velocity. Although the reason why deeper

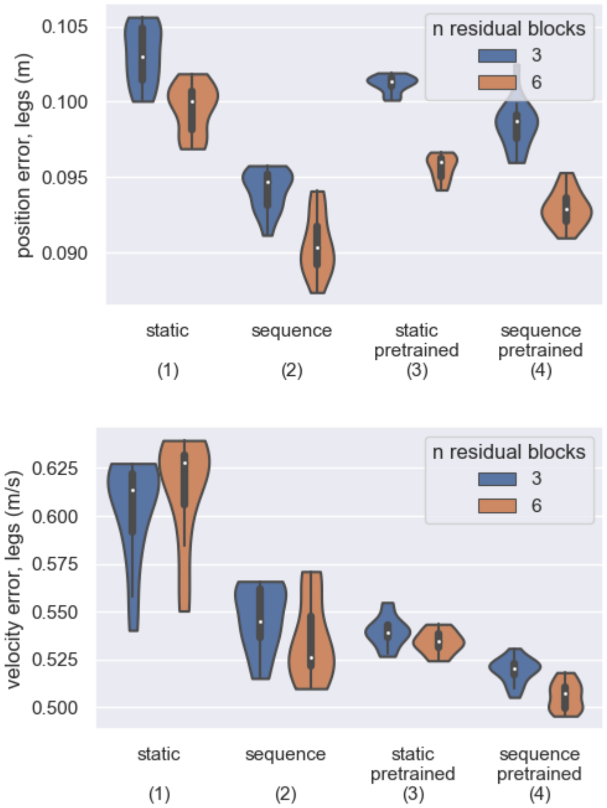

Figure 6. Average joint **position** (top) and **velocity** (bottom) errors on test sequences. Average over **legs only**.

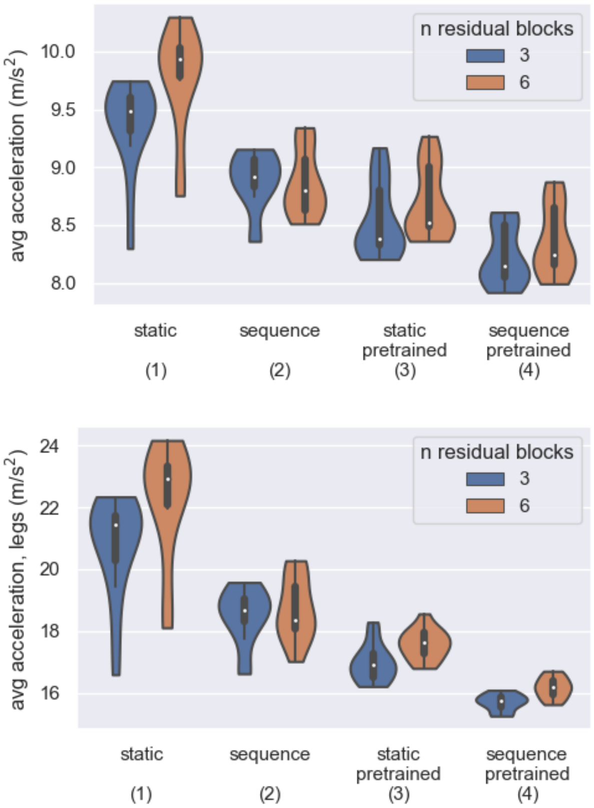

Figure 7. Average joint **accelerations** for different models on test sequences. Top: **full body**; bottom: **legs only**.

networks do not appear to achieve a lower velocity error is not obvious, one possible avenue to explore in future work could be to include a temporal consistency term in the loss.

5.4. Limitations

While our results are encouraging, we also observed several limitations with the proposed approach. Walking motion can be represented faithfully, but also fails in some instances to take full advantage of the temporal history in the way we might hope. Some limitations are inherent to our formulation of the problem: for example, neither the restriction to a single body shape nor the assumption that hand signals are always available would hold true for a real deployment. It would be interesting to explore these aspects in future work, in particular the even more extreme data imputation problem where a full body pose has to be computed in frames where only a head tracking signal is available.

6. Conclusion

We introduce a novel problem of generating plausible and diverse body poses based on an impoverished control signal coming from a head-mounted device. We show that, surprisingly, there is sufficient information in a stream of head and hand poses to reconstruct a plausible full body pose. To return to the problem described in the introduction, we recognize that the predicted pose may not be an *accurate* reconstruction of the person wearing an HMD; for lower body in particular, the signals are too limited for us to hope for that. However, remembering that our goal is to allow two or more people to communicate effectively and collaborate with each other, we believe that the system described here has valuable properties. First, for poses observed in the training data, the reconstructed head and hand positions are indeed a good match to the input signals, which is important as they form the primary communication cues. Second, the stream of predicted poses is surprisingly smooth, despite the fact that we've added no explicit regularization or latent space transition model to enforce this. Finally, the model learns to extract useful motion inference data from temporal history, and this is demonstrated quantitatively even in the very challenging evaluation of comparing to unseen ground truth motion. In summary, we believe that more important than accuracy is the question of *robustness*, and our results show that VAEs are a powerful tool in the context of predicting robust and plausible human motion, even from extremely limited input data.

References

[1] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1446–1455, 2015.

[2] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Lars Petersson, Stephen Gould, and Mathieu Salzmann. Contextually plausible and diverse 3D human motion prediction. *arXiv preprint arXiv:1912.08521*, 2019.

[3] 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 international conference on computer vision*, 2020.

[4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, page 408–416, New York, NY, USA, 2005. Association for Computing Machinery.

[5] Okan Arikan and David A Forsyth. Interactive motion generation from examples. *ACM Transactions on Graphics (TOG)*, 21(3):483–490, 2002.

[6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[7] Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Garrison W Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. *arXiv preprint arXiv:2003.04887*, 2020.

[8] Benjamin Biggs, Sébastien Ehrhadt, Hanbyul Joo, Benjamin Graham, Andrea Vedaldi, and David Novotny. 3d multibodies: Fitting sets of plausible 3d human models to ambiguous image data. *arXiv preprint arXiv:2011.00980*, 2020.

[9] 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 *European conference on computer vision*, pages 561–578. Springer, 2016.

[10] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6158–6166, 2017.

[11] Michael Büttner and Simon Clavet. Motion matching-the road to next gen animation. *Proc. of Nucl. ai*, 2015.

[12] Congqi Cao, Yifan Zhang, Yi Wu, Hanqing Lu, and Jian Cheng. Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3763–3771, 2017.

[13] Carnegie Mellon University. CMU MoCap Dataset.

[14] Alireza Fathi, Ali Farhadi, and James M Rehg. Understanding egocentric activities. In *2011 international conference on computer vision*, pages 407–414. IEEE, 2011.

[15] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

[16] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017.

[17] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 786–803, 2018.

[18] Félix G. Harvey and Christopher Pal. Recurrent transition networks for character locomotion. In *SIGGRAPH Asia 2018 Technical Briefs*, SA '18, New York, NY, USA, 2018. Association for Computing Machinery.

[19] Félix G. Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Trans. Graph.*, 39(4), July 2020.

[20] 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.

[21] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[22] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. Learned motion matching. *ACM Trans. Graph.*, 39(4), July 2020.

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

[24] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016.

[25] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-RNN: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5308–5317, 2016.

[26] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3509. IEEE, 2017.

[27] Brennan Jones, Yaying Zhang, Priscilla N. Y. Wong, and Sean Rintel. Belonging there: VROOM-ing into the uncanny valley of XR telepresence. In *CSCW 2021*. ACM, Association of Computing Machinery, November 2021.

[28] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *2020 International Conference on 3D Vision (3DV)*, pages 918–927. IEEE, 2020.

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[31] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

[32] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.

[33] Jogendra Nath Kundu, Maharshi Gor, and R. Venkatesh Babu. BiHMP-GAN: Bidirectional 3D Human Motion Prediction GAN. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8553–8560, 2019.

[34] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9887–9895, 2019.

[35] Robert LiKamWa, Zhen Wang, Aaron Carroll, Felix Xiaozhu Lin, and Lin Zhong. Draining our glass: An energy and heat characterization of Google Glass. In *Proceedings of 5th Asia-Pacific Workshop on Systems*, pages 1–7, 2014.

[36] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31:9605–9616, 2018.

[37] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015.

[38] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016.

[39] 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*, pages 5442–5451, Oct. 2019.

[40] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour. The KIT whole-body human motion database. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 329–336, July 2015.

[41] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019.

[42] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017.

[43] 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.

[44] Christos Mousas, Dimitris Anastasiou, and Ourania Spantidi. The effects of appearance and motion of virtual characters on emotional reactivity. *Computers in Human Behavior*, 86:99–108, 2018.

[45] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database HDM05. Technical Report CG-2007-2, Universität Bonn, June 2007.

[46] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. *CVPR*, 2020.

[47] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018.

[48] Frank C Park. Distance metrics on the rigid-body motions with applications to mechanism design. 1995.

[49] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019.

[50] Thammathip Piumsomboon, Gun A Lee, Jonathon D Hart, Barrett Ens, Robert W Lindeman, Bruce H Thomas, and Mark Billinghurst. Mini-me: An adaptive avatar for mixed reality remote collaboration. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages #46, 1–13, 2018.

[51] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[52] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016.

[53] Grégory Rogez, James S Supancic, and Deva Ramanan. First-person pose recognition using egocentric workspaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4325–4333, 2015.

[54] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1146–1161, 2019.

[55] Alla Safonova and Jessica K Hodgins. Construction and optimal search of interpolated motion graphs. In *ACM SIGGRAPH 2007 papers*, pages 106–es. 2007.

[56] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2325–2334, 2019.

[57] Hedvig Sidenbladh, Michael J Black, and David J Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *European conference on computer vision*, pages 702–718. Springer, 2000.

[58] Cristian Sminchisescu and Bill Triggs. Kinematic jump processes for monocular 3d human tracking. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003.

[59] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. 2017.

[60] Graham W Taylor and Geoffrey E Hinton. Factored conditional restricted Boltzmann machines for modeling motion style. In *Proceedings of the 26th annual international conference on machine learning*, pages 1025–1032, 2009.

[61] Graham W Taylor, Geoffrey E Hinton, and Sam T Roweis. Modeling human motion using binary latent variables. In *Advances in neural information processing systems*, pages 1345–1352. Citeseer, 2007.

[62] Franco Tecchia, Leila Alem, and Weidong Huang. 3D helping hands: A gesture based mr system for remote collaboration. In *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, VRCAI '12, pages 323—-328, New York, NY, USA, 2012. Association for Computing Machinery.

[63] Denis Tome, Thiemo Alldieck, Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernan Badino, and Fernando De la Torre. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *arXiv preprint arXiv:2011.01519*, 2020.

[64] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xR-EgoPose: Egocentric 3D human pose from an hmd camera. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7728–7738, 2019.

[65] Hsiao-Yu Fish Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. *arXiv preprint arXiv:1712.01337*, 2017.

[66] Dorin Ungureanu, Federica Bogo, Silvano Galliani, Pooja Sama, Xin Duan, Casey Meekhof, Jan Stühmer, Thomas J Cashman, Bugra Tekin, Johannes L Schönberger, et al. HoloLens 2 Research Mode as a tool for computer vision research. *arXiv preprint arXiv:2008.11239*, 2020.

[67] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7124–7133, 2019.

[68] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics*, 25(5):2093–2101, 2019.

[69] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 265–281, 2018.

[70] Haruka Yonemoto, Kazuhiko Murasaki, Tatsuya Osawa, Kyoko Sudo, Jun Shimamura, and Yukinobu Taniguchi. Egocentric articulated pose tracking for action recognition. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, pages 98–101. IEEE, 2015.

[71] Ye Yuan and Kris Kitani. 3d ego-pose estimation via imitation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 735–750, 2018.

[72] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time PD control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10082–10092, 2019.

[73] 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.

[74] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Thundr: Transformer-based 3d human reconstruction with markers. *arXiv preprint arXiv:2106.09336*, 2021.

[75] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.