FLAG: Flow-based 3D Avatar Generation from Sparse Observations

Sadegh Aliakbarian Pashmina Cameron Federica Bogo Andrew Fitzgibbon Thomas J. Cashman

Mixed Reality & AI Lab, Microsoft

https://microsoft.github.io/flag

Abstract

To represent people in mixed reality applications for collaboration and communication, we need to generate realistic and faithful avatar poses. However, the signal streams that can be applied for this task from head-mounted devices (HMDs) are typically limited to head pose and hand pose estimates. While these signals are valuable, they are an incomplete representation of the human body, making it challenging to generate a faithful full-body avatar. We address this challenge by developing a flow-based generative model of the 3D human body from sparse observations, wherein we learn not only a conditional distribution of 3D human pose, but also a probabilistic mapping from observations to the latent space from which we can generate a plausible pose along with uncertainty estimates for the joints. We show that our approach is not only a strong predictive model, but can also act as an efficient pose prior in different optimization settings where a good initial latent code plays a major role.

1. Introduction

Mixed reality technology provides new ways to interact with people, with applications in remote collaboration, virtual gatherings, gaming and education. *People* are at the heart of all these applications, and so generating realistic human representations with high fidelity is key to the user experience. Whilst external sensors and cameras [33] are effective, using only head-mounted devices (HMDs) to generate realistic and faithful human representations remains a challenging problem. The relevant data available from HMDs such as Microsoft HoloLens and Oculus Quest is limited to the location and orientation of the head and the location and orientation of the hands, obtained either via egocentric hand tracking [11, 39] or the signal from motion controllers. This is a very incomplete signal for the pose and motion of the full human body.

Although prior work has proposed human pose priors for generating 3D human body poses from partial and ambiguous observations such as images [3, 19, 42], 2D joints/keypoints [4, 26], and markers [9, 22, 44, 45], such observations are richer sources of data than those available in practice from HMDs. Despite the importance of this problem, there have been few attempts to generate full body pose from extremely sparse observations, i.e., head and hand position and orientation. Dittadi et al. [8] developed a variational autoencoder (VAE) to compress the head and hands inputs to a latent space, allowing a full-body pose to be generated by sampling from that latent space.

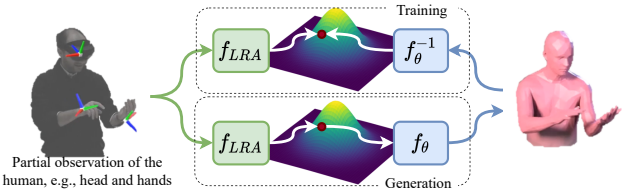


Figure 1. We generate a full body avatar given sparse HMD input (three SE3s for head and hands), by training a flow-based generative model that provides an invertible mapping between the base distribution and 3D human pose distribution. At test time, given the HMD signal, we predict a region in the latent space that is used as the input to the flow-based model to generate a pose.

We propose a new approach based on conditional normalizing flows for *sparse inputs*. Specifically, we learn the conditional distribution of the full body pose given the head and hand data via a flow-based model which enables an invertible mapping between the 3D pose distribution and the base distribution. Invertibility of our model then allows us to go further by learning a probabilistic mapping from the condition to the *high-likelihood* region in the same base distribution, as illustrated in Fig. 1. We name our approach a **fl**ow-based **a**vatar **g**enerative model (**FLAG**). The strengths of this design are: first, using a flow-based generative model enables *exact pose likelihood* computation in contrast to the approximate likelihoods seen in VAE-based pose priors [8, 26]. Second, the invertibility of our generative model allows us to compute the *oracle latent code*. During training, the oracle latent code then acts as the ground truth for our mapping function. This allows us to learn a representative mapping from the observed head and hands to the latent

space, making our approach a *strong predictive model*. Finally, when optimizing either in pose space or latent space, using our model as the pose prior provides a *superior initialization* in the latent space, making optimization very efficient.

2. Related Work

Several recent works generate 3D human body pose given partial observations, such as images [3, 18, 19, 35], 2D keypoints [4, 15, 19], HMDs [8], IMUs [10] and additional upper body tracking signals [41], or trajectories of partially visible body joints [16]. These methods usually require richer input than is available from commercial HMDs [40], whereas we wish to address the challenge of generating full body poses solely from HMD input. Most related work uses a generative model of human pose, either to directly predict the parameters of the body model [23, 25, 26] or as a pose prior in an optimization framework [4, 19, 45]. A few authors combine the two, either by training a network that mimics the behavior of an optimizer [15, 43] or an optimizer initialised using a neural network [18].

SMPLify [4] proposed a probabilistic 3D human pose prior based on a mixture of Gaussians. Pavlakos et al. [26] found SMPLify insufficiently expressive to model the complex human pose distribution and proposed VPoser, which uses variational autoencoders [17]. When using unconditional VAE-based pose priors, consistency with the observations has to be enforced via additional terms in the optimization cost function. In contrast, conditional VAE-based (CVAE) [38] pose priors use the observations to estimate the pose likelihood. Liang et al. [20] and Rempe et al. [28] use previously observed poses to condition the pose prior while Sharma et al. [36] use a CVAE to generate 3D human pose from 2D keypoints extracted from images.

Previous studies [6, 12, 27] have established that VAE-based approaches are challenging to train due to the heuristic nature of tuning the balance between the reconstruction and the KL divergence loss in the VAE's evidence lower bound (ELBO). If the goal is to learn a rich semantic latent space close to a normal distribution, the weight for the KL term needs to be relatively high (close to 1 as in the standard ELBO), which in turn leads to lower quality pose reconstruction through decoding. If one requires high-quality pose reconstruction, then the weights for the KL should be relatively small, e.g., 5×10^{-3} in VPoser [26], leading to a model that does not optimize the true ELBO with an imperfect latent representation. This push-pull effect on the weights of the two terms makes VAE training difficult, but it becomes even more challenging when strong conditioning signals, such as images, previous poses, or 2D keypoints [20, 28, 36] are introduced. If trained in the standard fashion, the conditioning signal is strong enough that the decoder can generate a pose given only the condition, and thus learns to ignore the latent variable [1, 2]. To avoid this, CVAE-based pose priors tend to assign a very small weight to the KL term, e.g., 4×10^{-4} as in Rempe et al. [28], to avoid posterior collapse [6, 12, 27].

Unlike VAE-based approaches, normalizing flow-based models represent the complex data distribution via a composition of invertible transforms and minimize the *exact* negative log-likelihood of the poses. Biggs et al. [3] use a flow-based model as a pose prior in an optimization problem where the goal is find a likely pose that minimizes the re-projection error given 2D keypoints. Zanfir et al. [42] use a flow-based pose prior to fit a 3D body model on 2D images in a weakly-supervised framework. Kolotouros et al. [19] extend these models to make them conditional on observed images enabling them to use a single model both as a pose prior and directly as a predictive model, allowing generation of a plausible 3D human pose given an image and a latent code ($z = \mathbf{0}$). These advances build confidence in highly expressive conditional flow-based models with rich conditional inputs such as images or keypoints.

Our approach In this paper, we push this line of investigation even further and propose FLAG, a conditional flow-based pose prior for *sparse inputs* which builds upon prior work by: (1) generating high-quality 3D poses from an extremely sparse conditioning signal, (2) providing *latent variable sampling*, by learning a mapping from the observation to the region in the latent space that generates a likely and plausible pose. This gives us a *strong predictive model* as well an *efficient pose prior* for optimization. Furthermore, we show that, in a conditional scenario, starting with $z = \mathbf{0}$ as in Kolotouros et al. [19] does not necessarily lead to the best predictive outcome and that our method provides a more promising alternative.

3. Preliminaries

Normalizing Flows. Normalizing flows [30] as likelihood-based generative models provide a path to an expressive probability distribution of the data. Unlike VAEs, where the main challenge is to find an appropriate approximate posterior distribution, normalizing flows only require a definition of a simple base distribution (also referred to as a prior) and a series of bijective transformations. These bijective transformations allow the model to map the data to the latent space and vice versa.

Given data $x \in \mathbb{R}^d$, the goal is to learn the joint distribution of data. Normalizing flows model x as a transformation T of a real vector $z \in \mathbb{R}^d$ sampled from the chosen base distribution $p_z(z)$, which could be as simple as a multivariate normal distribution. With invertible and differentiable T (and hence T^{-1}) and using the change of variable formula [32], we obtain the density of x as:

$$p_x(x) = p_z(z)\left|\det J_T(z)\right|^{-1} \tag{1}$$

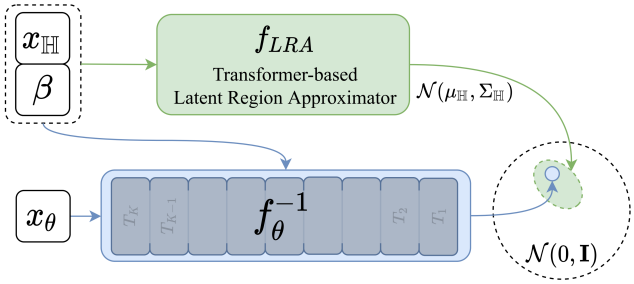


Figure 2. Overview of FLAG, consisting of a flow-based model f_θ and a latent region approximator f_{LRA}. During training f_θ aims to learn the distribution of x_θ and f_{LRA} aims to learn a mapping from the condition to the latent representation of x_θ. At test time, we sample a latent variable $z_{\mathbb{H}}$ via f_{LRA} and use that to generate a new pose via $\hat{x}_\theta = f_\theta(z_{\mathbb{H}}, [x_{\mathbb{H}}, \beta])$.

where J_T is the Jacobian of T. Since $z = T^{-1}(x)$, $p_x(x)$ can also be written in terms of x and the Jacobian of T^{-1}:

$$p_x(x) = p_z(T^{-1}(x)) \left| \det J_{T^{-1}}(x) \right| \tag{2}$$

Instead of one transformation, multiple simple transforms can be composed to form a complex transform $T = T_K \circ T_{K-1} \circ ... \circ T_1$ where T_i transforms z_{i-1} into z_i, z_0 is the latent variable in the base distribution and $x = z_K$. This composition can be built with neural networks that maximize the data log-likelihood. As shown in [30], $\log p(x)$ can be written as:

$$\log p(x) = \log p(z_0) - \sum_{i=1}^{K} \log \det \left| \frac{\partial T_i}{\partial z_i} \right| \tag{3}$$

SMPL Body Model. SMPL [23] is a parametric generative model of human body meshes. SMPL receives as input the 3D human poses in axis-angle representation θ and the body shape parameters β, and generates the body mesh represented as 3×6890 matrix $M = \texttt{SMPL}(\theta, \beta)$. With that, we define $\texttt{SMPL}(\theta, \beta).\texttt{HH()}$ to compute the head and hands location and orientation.

4. Proposed Method

We first define our problem statement, followed by an overview of our approach. We then describe the components of FLAG and the training and generation of full body poses.

4.1. Model Overview

Our task is to generate a full body pose x_θ given a sparse observation $x_{\mathbb{H}}$ and the shape parameters β. $x_\theta \in \mathbb{R}^{3\times J}$ represents joint rotations as axis-angle vectors for J body joints, and $x_{\mathbb{H}} \in \mathbb{R}^{9\times K}$ represents the global 6D joint rotation [46] and a 3D joint location for each of the $K = 3$ observations (head and hands). This information can be obtained from a parametric model of human body, e.g., SMPL [23].

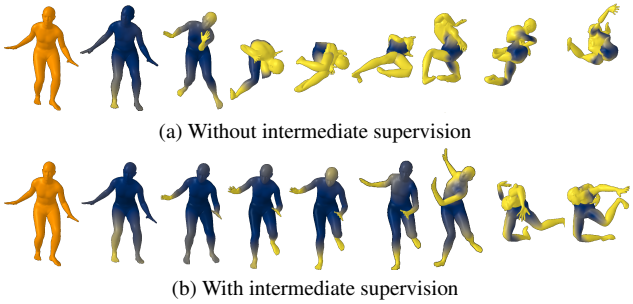
One valid way to generate x_θ from $x_{\mathbb{H}}$ is to learn the distribution of the body pose given the observed $x_{\mathbb{H}}$ and β via a conditional flow-based model f_θ. While this approach can effectively provide the likelihood of a given pose, the generation process remains incomplete; for generating a novel pose given $x_{\mathbb{H}}$ and β, one needs to sample a latent variable. However, the sampling process is completely independent of the observations. While [19] rely on the mean of the latent space $z = \mathbf{0}$ (a vector of all zeros) as the latent code to generate the full pose, we argue that there exists a latent code that represents x_θ better than $z = \mathbf{0}$. In fact, while $z = \mathbf{0}$ is the most likely latent code in the base distribution, it may not necessarily translate to the most likely pose in the pose space since there can be changes in the volume of the distribution (the second term in Eq. 3) through f_θ's transformations. To obtain such a latent code, our model estimates a sub-region in the normalizing flow base distribution, $\mathcal{N}(\mu_{\mathbb{H}}, \Sigma_{\mathbb{H}})$, given $x_{\mathbb{H}}$ and β, from which a latent variable can be sampled to generate the full body pose.

At test time, to generate a full body pose given $x_{\mathbb{H}}$ and β, we sample a latent code from $z_{\mathbb{H}} \sim \mathcal{N}(\mu_{\mathbb{H}}, \Sigma_{\mathbb{H}})$ and use that as an approximation of z_θ, the latent code that generates a full body pose. We use this latent estimate to generate a full body pose via $\hat{x_\theta} = f_\theta(z_{\mathbb{H}}, [x_{\mathbb{H}}, \beta])$. Next, we define f_θ and describe how we model $\mathcal{N}(\mu_{\mathbb{H}}, \Sigma_{\mathbb{H}})$.

4.2. Flow over Full Body Pose

We model the distribution of x_θ with a normalizing flow model. Our model f_θ is a conditional RealNVP [7] conditioned on $x_{\mathbb{H}}$ and β. This can be achieved by mapping x_θ from the pose distribution to the base distribution (and vice versa) via a composition of simple invertible transformations, where each transformation can stretch or shrink its input distribution.

While it is not straightforward to investigate the contribution of each invertible transformation in generating a human pose given z_θ sampled from the base distribution, we expect each successive transformation to add expressivity to the incoming distribution of human poses it acts upon. To intuitively understand the role of each transformation, we visualize how a human pose is formed through all transformations in a model. Fig. 3a illustrates how z_θ from the base distribution evolves through invertible transformations of f_θ, up to the last transformation which produces a pose from the pose distribution. As shown, most of the observable modifications to the intermediate distributions are happening in later stages, in which one observes a human-like pose being formed. We argue that this is because the only source of supervision is the ground truth (GT) pose that explicitly guides the last transformation block. To ease the training and get the most out of each transformation block in f_θ, we propose to introduce *intermediate supervision* to f_θ. In addition to having GT pose as the input to the last transformation



(a) Without intermediate supervision

(b) With intermediate supervision

Figure 3. Pose progression through transformations from the base distribution to the pose distribution ($T_K \leftarrow T_1$) with and without intermediate supervision. First column shows the GT pose. The poses are color-coded to show large errors from GT in yellow, with dark blue showing zero error.

block, we provide the GT pose as the input to the intermediate transformation blocks, as if they are the last block of a sub-network. This is possible because the transformations in f_θ do not modify the data dimension. As a result, intermediate transformation blocks are encouraged to produce reasonable human poses and their capacity is exercised fully. We illustrate the pose evolution through transformations with and without intermediate supervision in Fig. 3b) and also show in 2 that intermediate supervision leads to improved plausibility of the generated poses.

4.3. Latent Variable Sampling

To generate a novel pose given $x_{\mathbb{H}}$ and β, we need to sample a latent variable z from the base distribution and use that to generate a pose $\hat{x}_\theta = f_\theta(z, [x_{\mathbb{H}}, \beta])$. In a standard conditional flow-based model, one randomly samples $z \sim \mathcal{N}(0, \mathbf{I})$, hoping for a plausible pose to be generated by the model, or consider $z = \mathbf{0}$ [19]. Although these approaches yield valid solutions, we argue and empirically show that these do not constitute the best solution. This can be examined explicitly thanks to the invertibility of normalizing flows, where one can obtain the oracle latent code $z^* = f_\theta^{-1}(x_\theta, [x_{\mathbb{H}}, \beta])$. Since an oracle latent code is known during training, we train our model such that it learns to map the condition ($x_{\mathbb{H}}$ and β) into a region in the base distribution where z^* has a high likelihood. Utilizing z^* during training allows us to take into account the changes in the volume of the base distribution, and thus the changes in the probability mass around the latent code, when transformed from the base distribution to the pose space via f_θ. We model the region of interest with a Gaussian and learn its parameters $\mu_{\mathbb{H}}$ and $\Sigma_{\mathbb{H}} = \text{diag}(\sigma_{\mathbb{H}})^2$. Such a mapping should have two desirable properties: **(i)** It should be expressive, so that it can produce a representation of full body given the sparse observation. This is necessary to estimate a sub-region of the base distribution that represents the full body. **(ii)** It should account for uncertainty in human body representation given sparse observation. When only the head and hands are observed, there exist multiple plausible full-body poses. For each plausible pose, we need to know the corresponding sub-region in the base distribution. With these key properties in mind, we design a transformer-based mapping function with a discrete latent space.

Attention-based Latent Region Estimation. We propose a transformer-based model (with a transformer encoder) to model the mapping function, taking advantage of the self-attention mechanism which learns the relationships between different joints in the body during training. Briefly, the transformer encoder receives as input $x_{\mathbb{H}}$ and β, and estimates $\mathcal{N}(\mu_{\mathbb{H}}, \Sigma_{\mathbb{H}})$ wherein $\mu_{\mathbb{H}}$ is trained to be a good approximation of the oracle latent code z^*.

For such a distribution to be representative of the full body, we make several design choices to come up with the model illustrated in Fig. 4. First, training such a model using sparse inputs directly proves challenging. To make it easier for the model, we define an auxiliary task of generating x_θ from the output of the transformer encoder (via `ToPoseSpace` block in Fig. 4), initially aiming to reconstruct x_θ from full body joints and gradually decreasing the joint visibility (through masking) in the encoding until we provide only head and hands[1]. To further help the transformer learn the representation of the body, we introduce another auxiliary task of predicting the masked joints given the observed ones. Such gradual masking-and-prediction (`MaskedJointPredictor`) lets the model infer the full body representation through attention (layer) on the available joints in the input. To get a compact representation out of the transformer encoder, we apply pooling ($\texttt{Pool}_{\mathbb{H}}$) over output joints and take only the head and hand representation, as they are always unmasked.

Next, we make the output of the transformer encoder stochastic, to obtain uncertainty estimates for the predicted poses. We propose using a categorical latent space [13, 31, 34] over human pose from the output of the transformer encoder (via `ToLatentSpace`). One can sample a discrete latent variable from this distribution (via Gumbel-Softmax [13] for differentiablity) to generate x_θ with the defined auxiliary task or use the entire latent representation to estimate $\mathcal{N}(\mu_{\mathbb{H}}, \Sigma_{\mathbb{H}})$ (via `LatentRegionApproximator`) which contains information about a plausible pose and the associated uncertainty. To model the complex distribution of human motion efficiently, we need a relatively large number of latent categories. To remedy this, we use a 2D categorical latent space, as shown in Fig. 4. We model a G-dimensional latent variable each responsible for M modes, giving us a capacity to

[1]While in principle, masking can be done randomly, we follow the kinematic tree of the SMPL skeleton and start by masking the lower body joints, then the spine joints, followed by arm joints, and finally the pelvis (the root of the kinematic tree).

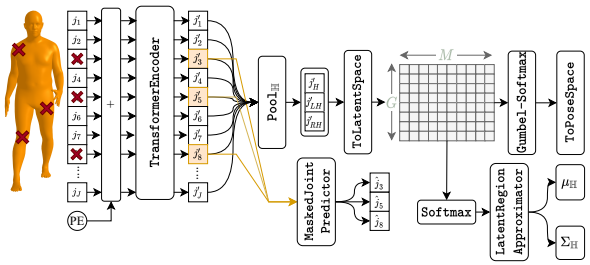


Figure 4. Transformer-based f_{LRA}. The attention-based encoder aims to learn the relationships between $x_{\mathbb{H}}$ and the rest of the body to generate an expressive representation of the body. A categorical latent space from the output of the transformer encoder allows us to predict a plausible pose and the associated uncertainty.

use M^G one-hot latent codes.

4.4. Learning

We use a dataset of diverse 3D human models, where each sample is a triplet $(x_\theta, x_{\mathbb{H}}, \beta)$, where β are the SMPL shape parameters. Our loss function $\mathcal{L}$ is given by

$$\mathcal{L} = \lambda_{\text{nll}}\mathcal{L}_{\text{nll}} + \lambda_{\text{mjp}}\mathcal{L}_{\text{mjp}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{lra}}\mathcal{L}_{\text{lra}} \tag{4}$$

where $\lambda_{.}$s are the weights associated with each term.

$\mathcal{L}_{\text{nll}}$: This term encourages the model to minimize the negative log-likelihood of x_θ under the model f_θ, following Eq. 3. Additionally, we take into account the log-likelihoods produced by the sub-networks of f_θ as the result of intermediate supervision discussed in Section. 4.2.

$$\mathcal{L}_{\text{nll}} = -\big(\log p_\theta(x_\theta) + \sum_{s\in S} w_s \log p^s_\theta(x_\theta)\big) \tag{5}$$

where S is the set of sub-networks of f_θ (e.g., from block T_i to T_1 for a pre-defined set of is), $p^s_\theta(x_\theta)$ is the likelihood of x_θ under sub-network s, and w_s is the weight associated to the sub-network that is proportional to the number of transformation blocks in each sub-network.

$\mathcal{L}_{\text{mjp}}$: To train the auxiliary task of masked joint prediction, we employ the term

$$\mathcal{L}_{\text{mjp}} = \sum_{j\in J_{\text{masked}}} \left\| \hat{x}^j_P - x^j_P \right\|^2_2 \tag{6}$$

where J_{masked} is the list of masked joints, x^j_P is the representation of the j^{th} joint in $\mathbb{R}^9$ (6D rotation and 3D location), and $\hat{x}^j_P$ is the corresponding prediction from the network.

$\mathcal{L}_{\text{rec}}$: This term acts on the output of the auxiliary task of decoding the full body pose from a discrete latent variable sampled transformer's categorical latent space, aiming to guide to build a meaningful discrete latent space.

$$\mathcal{L}_{\text{rec}} = \left\| \hat{x}^{\text{tps}}_\theta - x_\theta \right\|^2_2 \tag{7}$$

where $\hat{x}^{\text{tps}}_\theta$ is the output of `ToPoseSpace`.

$\mathcal{L}_{\text{lra}}$: Finally, this term encourages learning a Gaussian distribution $\mathcal{N}(\mu_{\mathbb{H}}, \Sigma_{\mathbb{H}})$ under which the oracle latent variable z^* has high likelihood.

$$\begin{aligned}\mathcal{L}_{\text{lra}} = &-\alpha_{\text{nll}} \log p_{\mathbb{H}}(z^*) + \alpha_{\text{rec}} \left\|\mu_{\mathbb{H}} - z^*\right\|^2_2 \\ &-\alpha_{\text{reg}}(1 + \ln \sigma_{\mathbb{H}} - \sigma_{\mathbb{H}})\end{aligned} \tag{8}$$

where $p_{\mathbb{H}}$ is the estimated sub-region of the base distribution. While the first term in Eq. 8 is enough to achieve this goal, we add the second term to implicitly encourage $\mu_{\mathbb{H}}$ to be similar to z^* and the third term discourages $\sigma_{\mathbb{H}}$ to be zero and thus avoids a deterministic mapping. Note that α_{reg} and α_{rec} can be relative small, but need to be present.

Although the entire model can be trained in an end-to-end fashion, we observed training f_θ first followed by training the latent region approximator is quite effective since we have access to a valid z^* from the beginning. The second training stage is quick, 4 GPU-hours. This two-stage training may also be useful in cases where one wishes to use a previously trained f_θ as a foundation model [5] and only train mapping functions for other data modalities, e.g., body markers or environment scans.

4.5. Conditional Generation

We can generate a full body pose given $x_{\mathbb{H}}$ and β by first computing $\mu_{\mathbb{H}}$ given the observation, then use $\mu_{\mathbb{H}}$ as an approximation of z_θ to generate a pose $\hat{x_\theta} = f_\theta(\mu_{\mathbb{H}}, [x_{\mathbb{H}}, \beta])$. To further enhance our quality of the generated pose, one can also use our flow-based model as a pose prior in optimization to minimize a cost function over the prior and the data. The optimization can be done either in pose space or in latent space. We use the LBFGS optimizer [21] throughout (see supp. mat. for further details).

Optimization in the pose space: The optimizer seeks a plausible human pose θ under our model that matches the observation $x_{\mathbb{H}}$. We optimize for θ by minimizing the cost:

$$\mathcal{C}(\theta) = -\log p_\theta(x_\theta) + ||\texttt{SMPL}(\theta, \beta).\texttt{HH()} - x_{\mathbb{H}}||^2 \tag{9}$$

Optimization in the latent space: The optimizer seeks a latent variable z that leads to a plausible pose under the model that matches the observation $x_{\mathbb{H}}$. Using generative functionality of the pose prior (f_θ) to generate a pose, we optimize for z by minimizing the cost:

$$\mathcal{C}(z) = -\log p(z) + ||\texttt{SMPL}(\hat{\theta}, \beta).\texttt{HH()} - x_{\mathbb{H}}||^2 + r \tag{10}$$

where $\log p(z)$ is the log-likelihood of the optimized z under the base distribution $\mathcal{N}(0, \text{I})$, $\hat{\theta} = f_\theta(z, [x_{\mathbb{H}}, \beta])$, and $r = ||z - \mu_{\mathbb{H}}||$ is a regularizer (see supplementary material) to implicitly prevent the latent code from straying too far from the initial guess (there is no signal for the lower body in the data term, i.e., the second term in Eq. 10).

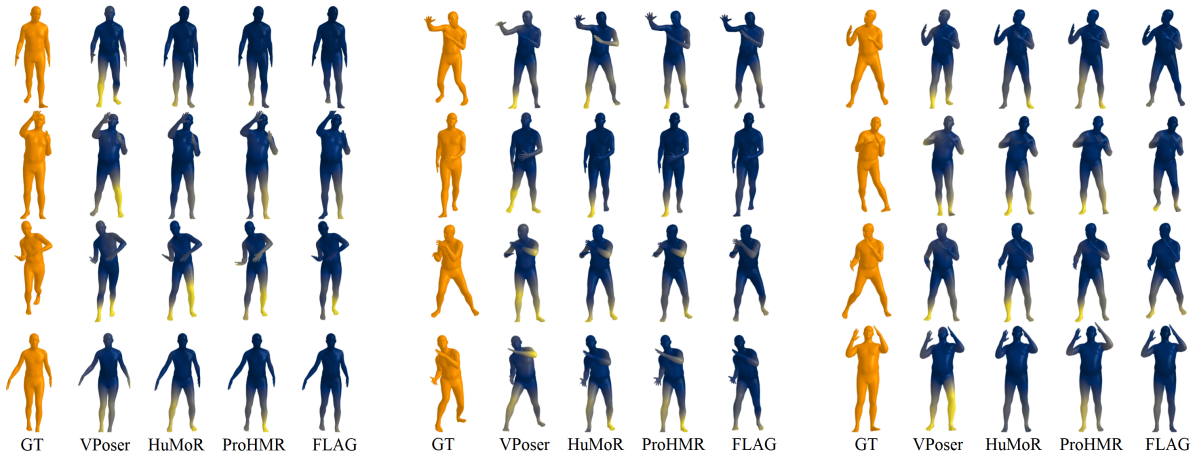


Figure 5. Qualitative results. First column (orange) shows the GT. Generated poses are color-coded to show large vertex errors in yellow.

Method	Upper Body MPJPE (↓)	Full Body MPJPE (↓)
VPoser-HMD	1.69 cm	6.74 cm
HuMoR-HMD	1.52 cm	5.50 cm
VAE-HMD	3.75 cm	7.45 cm
ProHMR-HMD	1.64 cm	5.22 cm
FLAG (Ours)	**1.29 cm**	**4.96 cm**

Table 1. Comparison of FLAG with existing methods on AMASS.

5. Experiments

We first introduce the dataset and then present the experimental results, ablation studies, and qualitative results of our approach (see supp. mat. for implementation details).

Dataset. We report results on AMASS [24], a large-scale motion capture dataset, with diverse poses represented with SMPL body model. We evaluate our approach and existing methods on the held out Transitions and HumanEVA [37] subset of the AMASS. The models are trained on the remaining datasets, excluding the dancing sequences [28].

Baselines. There have been a few efforts towards generating full body poses given head and hand inputs [8]. Our first baseline, which we call **VAE-HMD**, involves a two-step process. First a VAE encoder-decoder is trained on full body, without any condition. In the next step, another VAE is trained (starting from the frozen decoder) which encodes head and hand representation into the latent space and uses the previously trained full body decoder for generation. Since our approach is a conditional pose prior, we compare it with existing conditional pose priors after adapting them to our problem setting. ProHMR [19] is closest to our approach in terms of architecture since it is a conditional flow-based model. We adapt the conditioning signal to head and hands representation and include this as another baseline called **ProHMR-HMD**. Our third baseline is a conditional version of VPoser [26], a VAE-based approach, since it constitutes a strong and commonly used human pose prior in the literature, and refer to it as **VPoser-HMD**. We also evaluate another recently proposed CVAE-based pose prior, HuMoR [28], which learns a prior distribution given the conditioning signal. We adapt this approach to our scenario and refer to it as **HuMoR-HMD**. For all baselines we follow original implementation where available, otherwise follow the papers. We consider the same data and condition representations for all methods for a fair comparison. Following prevailing convention [8, 29], the avatar root is positioned at the origin.

Evaluation Metrics. To measure accuracy quantitatively, we report the mean per-joint position error (MPJPE) in cm. Since the quality of upper-body representation is of greater importance for AR, VR, and MR applications, we report the MPJPE of the upper body as well as that of the full body.

5.1. Comparison to Existing Approaches

We evaluate our approach in generating a plausible pose given sparse observations and compare it with existing methods. Table 1 summarizes this evaluation[2]. We do not use optimization for this comparison. Flow-based approaches, ProHMR-HMD and FLAG (Ours), generally yield lower full body error, but approaches that have conditional latent variable sampling tend to generate better upper bodies. This is the case with HuMoR and our approach, where the latent variable is sampled given head and hands information while for other techniques, a latent variable is sampled independent of the conditioning signal. The superiority of our approach is also evident in the qualitative results in Fig. 5, where FLAG yields least error compared to other techniques, with HuMoR producing relatively good upper body. We provide more qualitative results in the supplementary material.

5.2. Ablation Study

Effect of Intermediate Supervision. Building on Section 4.2, here we evaluate the effect of the proposed interme-

[2]The MPJPE for VAE-HMD on the standard AMASS test set is relatively high. We analyze this in the supplementary material, demonstrating that this is due to imperfect utilization of the latent space resulting from the two-stage training used in VAE-HMD approach.

Setting	Upper Body MPJPE (↓)	Full Body MPJPE (↓)
w/o Intermediate Sup.	1.64 cm	5.22 cm
w/ Intermediate Sup.	**1.39 cm**	**5.11 cm**

Table 2. Evaluating the effect of intermediate supervision.

Method	GT Pose	Manip. Pose (RD ↑)	Noise (RD ↑)
CVAE * (true ELBO)	29.68	29.68 (0.0)	32.40 (0.08)
VPoser-HMD *	34.79	35.56 (0.02)	$2.39\times1e3$ (0.98)
HuMoR-HMD *	46.02	49.21 (0.06)	$2.37\times1e4$ (0.99)
ProHMR-HMD †	110.72	282.01 (0.61)	$6.63\times1e7$ **(1.0)**
FLAG (Ours) †	98.54	489.66 **(0.80)**	$3.04\times1e13$ **(1.0)**

Table 3. Evaluating the generalizability of learned latent representation by examining the NLL of in- and out-of-distribution samples.* denotes VAE-based methods and †denotes NF-based methods.

Latent Variable Sampling	Upper Body MPJPE (↓)	Full Body MPJPE (↓)
Zeros ($z = \mathbf{0}$)	1.39 cm	5.11 cm
MLP ($z = \text{MLP}_{\mathbb{H}}$)	1.36 cm	5.05 cm
Ours ($z = \mu_{\mathbb{H}}$)	**1.29 cm**	**4.96 cm**

Table 4. Evaluating the effect of latent variable sampling. Comparing $z = \mathbf{0}$ [19], estimating z with an MLP, and our approach.

Latent Variable Sampling	Cosine Dist.(↓)	Sinkhorn Dist.(↓)
Random ($z \sim \mathcal{N}(0, \mathbf{I})$)	1.0	0.29
Zeros ($z = \mathbf{0}$)	1.0	0.22
Ours ($z = \mu_{\mathbb{H}}$)	**0.81**	**0.18**

Table 5. Distance to oracle latent code $z^* = f_\theta^{-1}(x_\theta)$.

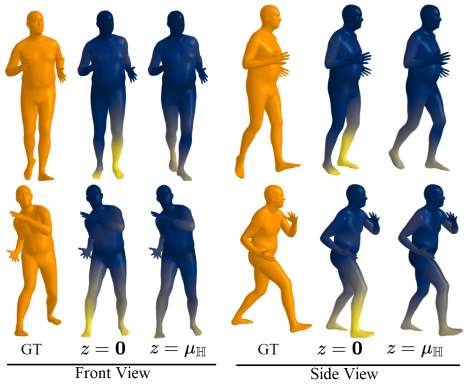


Figure 6. Qualitative evaluation of latent variable sampling, comparing our prediction from $z = \mathbf{0}$ and from $z = \mu_{\mathbb{H}}$. Generated poses are color-coded to show large vertex errors in yellow.

diate supervision. Such supervision provides an additional signal to intermediate transformation blocks of f_θ, allowing better convergence to a plausible pose throughout transformations when starting from the base distribution. This was visible in Fig. 3 and is also evident in our quantitative results in Table 2 wherein we show considerable improvement in the quality of generated poses.

Generalizability of Latent Representations. As various models are trained differently, with various training tricks such as KL term annealing or modifying the ELBO for mitigating posterior collapse, we define an auxiliary task to evaluate the quality of the learned latent space. To this end, we use the negative log-likelihood (NLL) metric to identify out-of-distribution (OOD) samples. We define in-distribution samples to be the poses from the ground truth test set, whereas the OOD samples are defined in two ways (1) manipulating ground truth poses by adding a small amount of noise to a subset of joints (2) creating pose-like random noise (random values within the range of natural poses). Table 3 summarizes how different models perform when detecting OOD samples. For a clearer comparison, we also report the relative difference (RD $= \frac{|\text{NLL}_{\text{OOD}} - \text{NLL}_{\text{GT}}|}{\max(\text{NLL}_{\text{OOD}}, \text{NLL}_{\text{GT}})}$) between the NLL of the models for OOD samples and that of the ground truth poses, which higher is better. It can be seen that flow-based models are typically better at detecting OOD samples, demonstrating a richer learned latent representation, whereas VAE-based ones are less effective despite utilizing various techniques to avoid posterior collapse. For reference, we also provide the results of a CVAE trained with true ELBO.

Effect of Initial Latent Code. A key contribution of this work is the probabilistic mapping from the condition to a sub-region in the latent space that leads to a highly plausible pose. In Table 4, we compare our approach, $z = \mu_{\mathbb{H}}$ with the proposal of ProHMR-HMD [19] which claims $z = \mathbf{0}$ yields the most plausible pose. While $z = \mathbf{0}$ yields *a* plausible pose, this experiment shows the existence of a better latent code, $z_{\mathbb{H}}$ that leads to a more plausible pose that has a high likelihood under the model. This is also shown in Table 5, where we compute the distance between the oracle latent code $z^* = f_\theta^{-1}(x_\theta, [x_{\mathbb{H}}, \beta])$ to the latent code from our approach as well as that of [19]. For the sake of completeness, we also compare our method with an MLP that learns to find a good latent code given the condition, which we refer to as $z = \text{MLP}_{\mathbb{H}}$ in Table 4. In addition to quantitative evaluation, Fig. 6 shows the effect of a proper initial latent code in generating pose from sparse observation.

We also observed that initial latent variable affects the quality of predicted poses refined via optimization in either the pose space or the latent space, as described in Section 4.5. We evaluate this in Fig. 7, where we use flow-based approaches as a pose prior in the optimization process and report the MPJPE. Consistent with Table 4, the results demonstrate that a proper initialization leads to a better performance. Given a fixed optimization budget, our method reaches a desired error threshold quicker owing to (a) a better initialization and (b) more reliable likelihood estimates (supported by results in 3). For instance, even after 50 iterations of optimization, ProHMR-HMD [19] does not outperform the solution reached by our approach after 2 optimization iterations regardless of the (pose or latent) space we optimize in. Finally, we also demonstrate that optimization in the latent space generally yields lower error compared to optimization in the pose space, for either model designs.

Partial Hand Visibility. All methods presented assume that

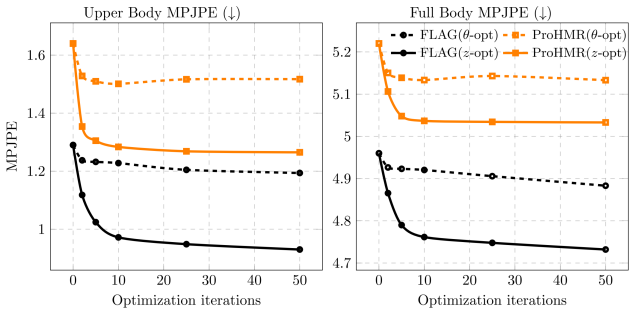


Figure 7. MPJPE as a function of optimization iteration. As shown here, ProHMR-HMD requires 50 iterations of optimization in the latent space to achieve an error on par with our approach with a proper initial latent code $z = \mu_{\mathbb{H}}$ without any optimization. Optimizing in the latent space yields lower error compared to optimizing in the pose space.

head and hand signals are always visible. In practice, one or both hands may go out of the field of view (FoV); real-world systems need to be robust to this. To make our model robust to hands going out of FoV, we fine-tune our model with random hand masking ($p = 0.2$) for 10 epochs. The use of progressive joint masking enables us to use fine-tuning for this purpose. In Fig. 8, we demonstrate that FLAG can generate highly plausible poses under partial or no hand observations[3].

Limitations and Future work. While FLAG is capable of generating highly plausible poses given extremely sparse observations in the majority of scenarios, it may fail to generate complex, less common lower-body poses, e.g., martial arts (examples are provided in the supp. mat.), potentially because these poses are not very common in the training dataset. FLAG uses only static pose information; extending FLAG to consume temporal data is a natural research direction. We use only HMD signal as the input to the model, whereas in some AR/VR scenarios, other modalities such as audio or environment scans may also be available. Although FLAG aims to find a better latent code to generate a plausible pose, there may still be a considerable gap between our estimated latent code and the oracle one (see Table 5). Further exploration in this area may lead to more faithful and accurate avatar poses.

Societal impact. While current datasets such as AMASS have a large number of poses, the data comes from 346 subjects who may not represent the true diversity of the global population. We have more work to do as a community to represent people of all age groups, and people with disabilities (*e.g.* wheelchair users, amputees). For anyone with a body morphology outside of the distribution represented in the datasets, we should ask: 1) Does the technology work for them? 2) Can they choose how they want to be represented?

[3]To get the uncertainty maps in Fig. 8, we generate K poses from $z \sim \mathcal{N}(\mu_{\mathbb{H}}, \Sigma_{\mathbb{H}})$ and compute the vertices' distance of these sampled poses from the one generated with $z = \mu_{\mathbb{H}}$.

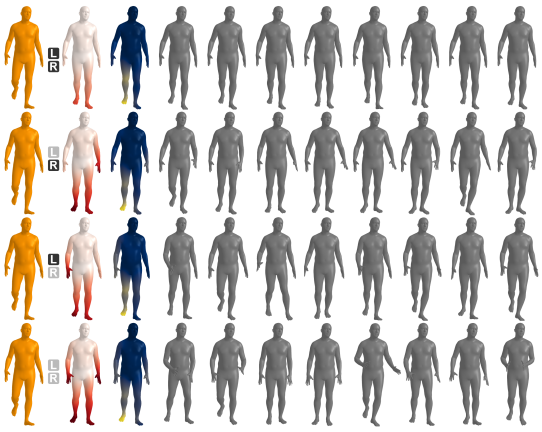


Figure 8. Qualitative results of FLAG when dealing with partially visible hands. From left to right, we illustrate the GT, the avatar's hand visibility status (black box is visible, gray box is invisible), uncertainty map on the pose from $z_{\mathbb{H}} = \mu_{\mathbb{H}}$ colorized based on the uncertainty (white is certain, red is uncertain), the pose from $\mu_{\mathbb{H}}$, followed by generated poses starting with samples from $\mathcal{N}(\mu_{\mathbb{H}}, \Sigma_{\mathbb{H}})$.

There could be negative outcomes from representing an individual in a way that removes a disability from view, for example. Mixed reality applications bring the promise of enhanced remote collaboration and communication, but there may also be potential negative societal impacts: misrepresentation including impersonation, further marginalization of socio-economically disadvantaged groups caused unknowingly or intentionally. Even so, with mindful deployment of technology and appropriate governance, we remain positive that realistic human representations can help the world grow closer without the harmful environmental impact of long-distance travel.

6. Conclusion

We presented FLAG, a new approach to generate plausible full body human poses from sparse HMD signals. FLAG is a conditional flow-based generative model of the 3D human body from sparse observations; we not only learn a conditional distribution of 3D human body, but also a probabilistic mapping from the observation to the latent space from which we generate plausible poses with uncertainty estimates. We show that our approach is both a strong predictive model, and an efficient pose prior in different optimization settings, thanks to our latent variable sampling mechanism. Experimental evaluation and ablation studies demonstrated that our method outperforms state of the art methods on the challenging AMASS dataset, requires fewer optimization iterations and leads to a very low error.

Acknowledgements. We thank Tom Minka and Darren Cosker for comments that greatly improved the manuscript. Pose visualizations were produced using ScenePic [14].

References

[1] Sadegh Aliakbarian, Fatemeh Saleh, Lars Petersson, Stephen Gould, and Mathieu Salzmann. Contextually plausible and diverse 3D human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11333–11342, 2021. 2

[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. 2

[3] Benjamin Biggs, Sébastien Ehrhadt, Hanbyul Joo, Benjamin Graham, Andrea Vedaldi, and David Novotny. 3D multi-bodies: Fitting sets of plausible 3D human models to ambiguous image data. 2020. 1, 2

[4] 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. 1, 2

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 6

[6] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *SIGNLL Conference on Computational Natural Language Learning (CONLL)*, 2016. 2

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. *International Conference on Learning Representations, ICLR*, 2017. 4

[8] Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lundell, Thomas J Cashman, and Jamie Shotton. Full-body motion from a single head-mounted device: Generating SMPL poses from partial observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11687–11697, 2021. 1, 2, 6, 7

[9] Nima Ghorbani and Michael J Black. Soma: Solving optical marker-based mocap automatically. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11117–11126, 2021. 1

[10] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human POSEitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4318–4329, 2021. 2

[11] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (TOG)*, 39(4):87–1, 2020. 1

[12] 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. 2016. 2

[13] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. *International Conference on Learning Representations, ICLR*, 2017. 5

[14] Matthew Johnson and Jamie Shotton. ScenePic: A lightweight and zero-install 3D visualization library. 2021. 10

[15] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2

[16] Kacper Kania, Marek Kowalski, and Tomasz Trzciński. TrajeVAE–controllable human motion generation from trajectories. *arXiv preprint arXiv:2104.00351*, 2021. 2

[17] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *International Conference on Learning Representations, ICLR*, 2014. 2

[18] 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 of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 2

[19] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11605–11614, 2021. 1, 2, 3, 4, 6, 8, 9

[20] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion VAEs. *ACM Transactions on Graphics (TOG)*, 39(4):40–1, 2020. 2

[21] Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989. 6

[22] Matthew Loper, Naureen Mahmood, and Michael J Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):1–13, 2014. 1

[23] 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. 2, 3

[24] 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. 6

[25] Ahmed AA Osman, Timo Bolkart, and Michael J Black. STAR: Sparse trained articulated human body regressor. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 598–613. Springer, 2020. 2

[26] 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. 1, 2, 6

[27] Ali Razavi, Aäron van den Oord, Ben Poole, and Oriol Vinyals. Preventing posterior collapse with delta-VAEs. *International Conference on Learning Representations, ICLR*, 2019. 2

[28] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3D human motion model for robust pose estimation. *International Conference on Computer Vision*, 2021. 2, 6

[29] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *European Conference on Computer Vision*, pages 71–87. Springer, 2020. 7

[30] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 2, 3

[31] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. Meshtalk: 3D face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1173–1182, October 2021. 5

[32] Walter Rudin. *Real and Complex Analysis, 3rd Ed.* McGraw-Hill, Inc., USA, 1987. 3

[33] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 1

[34] Fatemeh Saleh, Sadegh Aliakbarian, Hamid Rezatofighi, Mathieu Salzmann, and Stephen Gould. Probabilistic tracklet scoring and inpainting for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14329–14339, 2021. 5

[35] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16094–16104, 2021. 2

[36] 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. 2

[37] 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

[38] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491, 2015. 2

[39] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 1

[40] Dorin Ungureanu, Federica Bogo, Silvano Galliani, Pooja Sama, Casey Meekhof, Jan Stühmer, Thomas J Cashman, Bugra Tekin, Johannes L Schönberger, Pawel Olszta, et al. Hololens 2 research mode as a tool for computer vision research. *arXiv preprint arXiv:2008.11239*, 2020. 2

[41] Dongseok Yang, Doyeon Kim, and Sung-Hee Lee. LoBSTr: Real-time lower-body pose prediction from sparse upper-body tracking signals. In *Computer Graphics Forum*, volume 40, pages 265–275. Wiley Online Library, 2021. 2

[42] 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. 1, 2

[43] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3D human pose and shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14484–14493, 2021. 2

[44] 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. 1

[45] 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. 1, 2

[46] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 3