PhaseMP: Robust 3D Pose Estimation via Phase-conditioned Human Motion Prior

Mingyi Shi
The University of Hong Kong
myshi@cs.hku.hk

Sebastian Starke
Meta
sebastian.starke@mail.de

Yuting Ye
Meta
yuting.ye@gmail.com

Taku Komura
The University of Hong Kong
taku@cs.hku.hk

Jungdam Won*
Seoul National University
jungdam@imo.snu.ac.kr

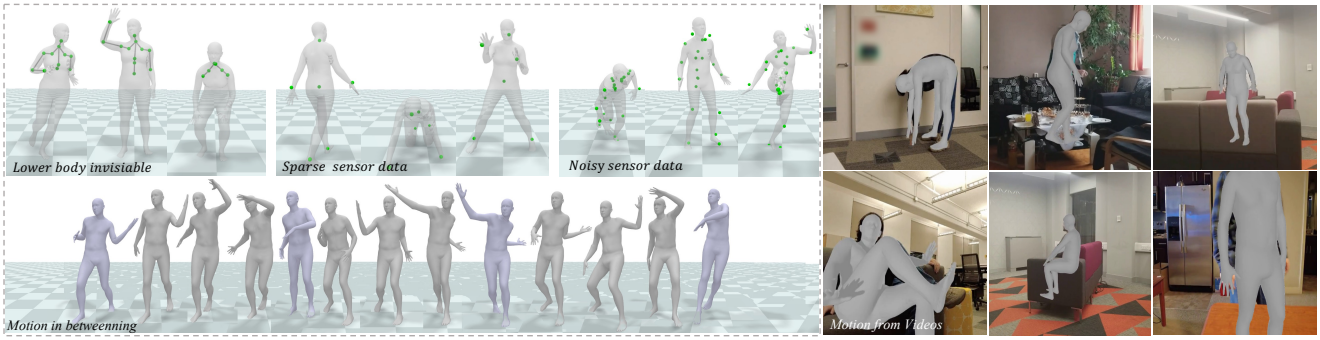

Figure 1. Our novel human motion prior *PhaseMP* enables us to further constrain the predictions in both nominal and challenging scenarios, resulting in more natural and stable movements. The figure demonstrates the generated motions in several challenging scenarios, which involve generation from incomplete observation in spatial or temporal domain (left), as well as generation from raw videos where heavy occlusion exists (right).

Abstract

*We present a novel motion prior, called **PhaseMP**, modeling a probability distribution on pose transitions conditioned by a frequency domain feature extracted from a periodic autoencoder. The phase feature further enforces the pose transitions to be unidirectional (i.e. no backward movement in time), from which more stable and natural motions can be generated. Specifically, our motion prior can be useful for accurately estimating 3D human motions in the presence of challenging input data, including long periods of spatial and temporal occlusion, as well as noisy sensor measurements. Through a comprehensive evaluation, we demonstrate the efficacy of our novel motion prior, showcasing its superiority over existing state-of-the-art methods by a significant margin across various applications, including video-to-motion and motion estimation from sparse sensor data, and etc.*

*Corresponding Author

1. Introduction

Estimating human poses and motions from real-world observations is a fundamental problem in computer vision with numerous potential applications, including human motion understanding, surveillance, motion capture, and human-computer interaction. Recently, there have been significant improvements in 3D human pose/motion estimation from 2D image/video via deep learning; specifically by deploying a data-driven system trained on a paired dataset [23, 31] to learn a mapping from 2D input images to 3D body motions [42, 47, 25, 57, 66]. However, these algorithms still struggle in challenging scenarios given unconstrained real-world data. For example, highly dynamic movements can cause motion blurs, and the blurry appearance of the body and the environment can lead to failures in 2D key-point detection, which is a preprocessing step widely used in many 3D pose estimation algorithms. Moreover, occlusions or a restricted camera field-of-view can result in partial or complete lack of pose information. Al-

though certain approaches have attempted to address this challenge by incorporating temporal consistency or formulating it as a denoising problem in the presence of artificial noise [9, 31, 50], they often struggle when dealing with long-period occlusion lasting beyond one second.

In this paper, we aim to develop a 3D pose/motion estimation algorithm as shown in Fig 1 that can work robustly in challenging scenarios illustrated above. More specifically, we propose a novel motion prior, called *PhaseMP*, modeling the probability distribution on pose transitions. Our novel motion prior draws inspiration from previous models [49, 38]; however, the key distinction lies in its combination with a frequency domain feature extracted from a periodic autoencoder [56], further enhancing the quality and robustness of 3D pose estimation. Intuitively speaking, the phase features play a role akin to long-term physical momentum, facilitating the generation of smooth and stable movements by enforcing a unidirectional motion (i.e., no backward movement in time). Consequently, phase features contribute to imbuing the resulting motions with greater naturalness, particularly in scenarios where the laws of physics play a significant role. This characteristic extends to challenging scenarios where other input features (e.g., joint location) are only partially observable. The underlying ambiguity arising from insufficient contextual cues can be significantly reduced by enforcing the model to maintain the current momentum.

We demonstrate the effectiveness of our motion prior in solving challenging video-to-motion estimation problems, where a large portion of the body is partially invisible for a long period of time due to occlusions or being out-of-sight. Additionally, we showcase its capabilities in addressing other problems such as denoising 3D motion data or estimating full 3D pose from end-effectors only. Through comprehensive evaluation, we demonstrate that our system outperforms state-of-the-art baselines with a large margin.

The paper's contributions can be summarized as follows: (1) a novel motion prior, combined with the phase feature, applicable to a wide range of 3D human pose estimation systems, (2) a new optimization framework incorporating phase feature energy, which can work robustly for many challenging scenarios where the observation is incomplete or ambiguous in temporal and spatial domains, (3) a comprehensive evaluation showcasing that our system outperforms existing state-of-the-art methods by a large margin, not only in ordinary scenarios but also in challenging scenarios.

2. Related Work

3D Human Pose Estimation As a tool in understanding the human-centered world, there has been a significant body of work focusing on estimating 3D human pose, which is typically represented by a positional skeleton [42, 20, 1, 47, 52, 10, 22, 24, 75], or parametric mesh model [4, 25, 46, 26, 33, 31, 15], from various real-world observations. These approaches can be categorized into two groups: optimization-based and learning-based methods.

The optimization-based methods typically rely on a known transformation function f (e.g., 3D-to-2D projection), and the 3D poses, as the optimization target x, are obtained by optimizing an objective function that encourages the transformation $f(x)$ to be close to the observations, represented in the form of 2D joint positions [13, 2] or body mask images [35]. In SMPLify [4], the optimization can be accelerated by constraining the body state using a linear model SMPL [40], the realism of the optimized poses is further improved by incorporating a learned pose prior [46].

However, accurately modeling real-world transformations [32, 65] can be challenging, and test-time optimization is also computationally expensive. To address these limitations, learning-based methods have emerged as a popular alternative, leveraging data to simplify the pose estimation process. These methods typically require a paired dataset [23, 64, 31] of 3D poses and corresponding observations, and learn an end-to-end mapping from the observations to the 3D poses using different specifications, such as known-camera-projection [42, 47], weak-perspective assumption [25, 48, 58, 59], or kinematic structure [52, 36]. However, these approaches often struggle in accurately predicting some key parameters in the wild environment, and may require significant amounts of supervision [34], which can be constrained by the dataset and network architecture.

In our project, we employ an optimization-based approach to achieve more robust pose estimation in challenging scenarios, such as long-term occlusion, where existing methods still struggle to perform effectively.

3D Human Motion Estimation In addition to estimating pose independently on a frame-by-frame basis, existing methods leverage temporal information to improve estimation accuracy. One popular approach is to encode the pose sequence using neural differential operators. For instance, [47] shows that a dilated temporal convolutional network with a sizeable receptive field significantly outperforms single frame methods. More advanced context-aware sequential encoders, such as graph convolutions [30, 75, 6, 78, 22] and transformers [39, 17, 76, 37, 41], which are commonly used in neutral language processing for encoding the long sequence data with parallel multi-head attention mechanism, have also been developed for this task. Additionally, the fusion of spatial and temporal information [71, 51, 77, 6] can be used to avoid conflicts in the sequential model and improve estimation performance. Moreover, adversarial training [31] between real human motions and predicted motion is another option to enhance the realism of pose sequences.

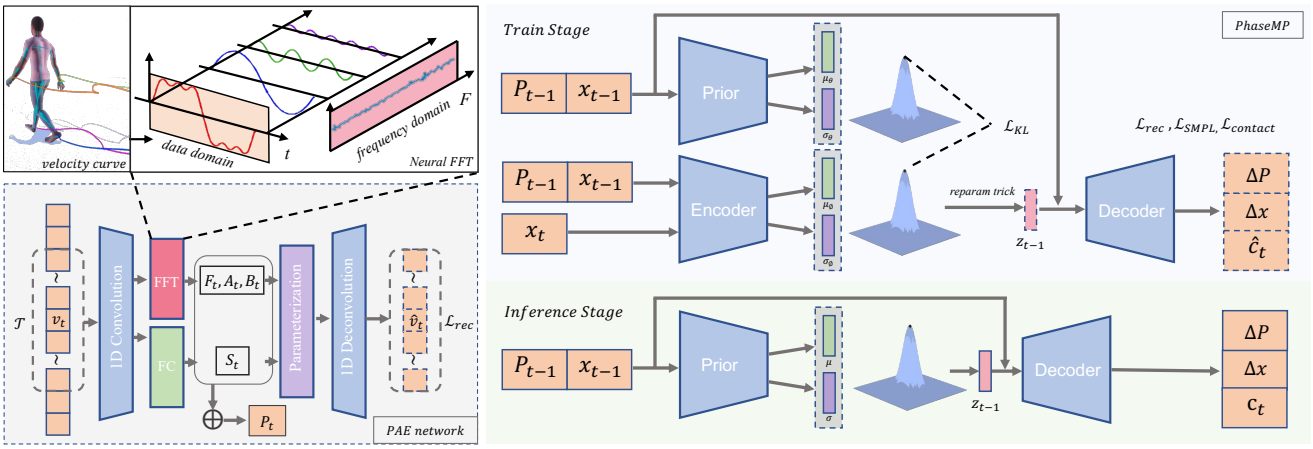

Figure 2. The overview of our system. The left figure illustrates the phase feature extraction process, which takes a window of joint velocities with the duration of $\mathcal{T}$ as input, then produces frequency domain periodic parameters $[F_t, A_t, B_t, S_t]$ as output, from which the phase feature P_t is computed by updating these parameters cyclically. The right figure is our phase-conditioned human motion prior, called *PhaseMP*, based on the structure of conditional VAE. It consists of a *Prior* network, an *Encoder* network, and a *Decoder* network. All the networks are used during the training stage, while only the *Prior* and the *Decoder* are used during the inference stage.

Another line of work tackles sequential pose estimation with learned powerful generative priors, which incorporate probabilistic models that can capture the nature of human movements; such works are based on mixtures-of-Gaussians [21], pose embeddings [45, 12, 63], neural distance field [61], variational autoencoder [14, 38, 19, 67], and the diffusion model [11, 68]. Given observations, the motion estimation is performed by searching for the most plausible alignment on the motion manifold. HuMoR [49] achieves impressive motion generation by modeling the pose transitions instead of modeling the poses directly. Their system can support various types of inputs such as RGB-D videos and 2D/3D joint position sequences. Our system also models the pose transitions but with a particular emphasis on incorporating a frequency domain feature for enhancing the robustness of pose estimation in challenging scenarios.

Measuring physical realism or consistency can also lead to a significant improvement in the accuracy of the predicted pose. For example, a consistent skeleton [60, 52, 8] or physics-inspired metrics [44, 16, 53, 69] such as foot sliding, foot-floor penetration, human-environment interaction [72, 74, 73] can improve temporal coherence and reduce the searching space in generating motions.

Motion Frequency Feature The conversion of motion signals from the time domain to the frequency domain [18] has been utilized for motion editing [5, 27], stylization [62, 70], and compression [3]. Compared to the original data domain, the features in frequency domain often remains consistent over a longer period of time, or between different motion clips, making it useful for synthesizing high quality transition motion [56]. We leverage frequency domain features extracted from a large-scale motion database to accurately estimate human movements in videos, including those frames in which the body is partially or fully occluded.

3. System Overview

Out system predicts the full 3D human motion sequence from incomplete data, such as 2D body landmark sequences extracted from videos, 3D end-effector sequences, or noisy 3D joint position sequences. Our system is composed of the periodic autoencoder to extract the phase feature, the phase-conditioned motion prior, and the run-time optimization module. Given a large motion database, the periodic autoencoder is first trained (see Sec 4.1). Using the extracted phase features as additional inputs, a motion prior based is trained (see Sec 4.2, 4.3). At inference time, given an input observation sequence, the 3D human motion is computed via optimization with energy functions that ensure both accuracy and realism (see Sec 5).

4. Phase-conditioned Motion Prior

We first describe how we compute the phase features using a periodic autoencoder trained with a large motion database. Then, we explain the structure of phase-conditioned motion prior and the loss functions to train it.

4.1. Deep Phase Feature

The periodic autoencoder (PAE) [56], which we use to compute the phase features, is depicted on the left side of Figure 2. It is an autoencoder equipped with convolution

and FFT (fast Fourier transform) layers in its intermediate structure, allowing it to compute embeddings in the frequency domain given joint velocities as inputs. Intuitively speaking, it learns alignments of periodic signals existing in motions, the learned embedding can play a role of physical momentum as a result. More specifically, at frame t, PAE uses a window of 3D joint velocities $X_t \in \mathbb{R}^{3\times J\times N}$ as input, where J, N represent the number of joints and the size of window, respectively. Given X_t for the frame t, the encoding process includes a sequence of 1D convolutions followed by a differentiable FFT layer:

$$\boldsymbol{A}_t, \boldsymbol{B}_t, \boldsymbol{F}_t = \mathrm{FFT}(\mathrm{Conv}(X_t)), \tag{1}$$

where $\boldsymbol{A}_t$, $\boldsymbol{B}_t$, and $\boldsymbol{F}_t$ are amplitudes, offsets, and frequencies, and phase shift $\boldsymbol{S}_t$ of periodic embeddings are obtained by a separate fully-connected network:

$$(s_x, s_y) = FC(\mathrm{Conv}(X_t)), \boldsymbol{S}_t = atan2(s_y, s_x), \tag{2}$$

where $atan2$ is a 2-argument arc-tangent. Then $\boldsymbol{A}_t, \boldsymbol{B}_t, \boldsymbol{F}_t$, $\boldsymbol{S}_t$ are used to compose phase features in the decoding process, which is conducted by first reconstructing the feature maps F_t in the temporal domain:

$$F_t = \boldsymbol{A}_t \cdot sin(2\pi \cdot (\boldsymbol{F}_t \cdot \mathcal{T} - \boldsymbol{S}_t)) + \boldsymbol{B}_t \tag{3}$$

where $\mathcal{T}$ is a known time window. This is followed by a 1D deconvolution $X'_t = \mathrm{DeConv}(F_t)$ for reconstructing the original signals. The entire PAE is trained using the reconstruction loss:

$$\mathcal{L}_{PAE} = MSE(X_t, X'_t) \tag{4}$$

Once PAE is trained, the deep phase feature P_t are computed as follows using the frequency domain parameters:

$$P_t = [\boldsymbol{p_t}, \boldsymbol{F}_t, \boldsymbol{A}_t]. \tag{5}$$

where $\boldsymbol{p_t} = (\boldsymbol{A}_t \cdot \sin(2\pi \cdot \boldsymbol{S}_t), \boldsymbol{A}_t \cdot \cos(2\pi \cdot \boldsymbol{S}_t))$ is called the phase manifold vectors that periodically change over time, and the last two variables are frequency and amplitude.

4.2. Motion Prior Modeling

The right side of Figure 2 illustrates our novel phase-conditioned motion prior. The structure is inspired by MotionVAE [38] and HuMoR [49]; both models, including ours, are based on conditional VAEs [29] where the distribution of plausible pose transition is learned. The key differences are that our model is conditioned by the deep phase feature P_t extracted from pre-trained periodic autoencoder [56] in addition to the pose feature, and sinusoidal activation layers [54] are also incorporated to fully utilize the periodic nature of our phase feature.

For the pose feature, we use the same representation used in HuMoR [49], which includes position, orientation, and corresponding velocities for all the joints. Our motion prior receives the previous phase feature P_{t-1}, the previous pose feature x_{t-1}, and the next pose feature x_t as input, then predicts the change of pose Δx and phase ΔP features as the output, from which the features are updated as follows:

$$\begin{aligned} \hat{x}_t &= x_{t-1} \bigoplus \Delta x \\ \hat{P}_t &= P_{t-1} + \Delta P, \end{aligned} \tag{6}$$

where $\bigoplus$ is a differentiable integration operator to compute the current pose given the previous pose and its change Δx, where positional and rotational components are updated via addition and multiplication, respectively.

In this process, ΔP and Δx are computed from a decoder $\mathcal{D}$ conditioned on a latent variable (i.e. embedding) z that describes the possible pose transition. We use a learnable prior $\mathcal{R}$ and an encoder $\mathcal{E}$ (i.e., posterior) similarly to HuMoR [49]. They are defined as follows:

$$\begin{aligned} z'_{t-1} &= \epsilon(\mathcal{R}(x_{t-1}, P_{t-1})) \\ z_{t-1} &= \epsilon(\mathcal{E}(x_{t-1}, x_t, P_{t-1})) \\ \Delta x, \Delta P, c_t &= \mathcal{D}(x_{t-1}, z_{t-1}, P_{t-1}), \end{aligned} \tag{7}$$

where ϵ denotes a re-parameterization operation [28], z'_{t-1} and z_{t-1} represents the latent variables sampled from the Gaussian distribution, generated from the prior and the encoder, respectively, and c_t is the contact label to further enhance motion quality.

4.3. Prior Training

The pose in the next frame is predicted by sampling a Gaussian distribution produced by the prior model. In the training stage, following the CVAE [55], as shown in the right side of Figure 2, the prior model $\mathcal{R}$ is trained with posterior $\mathcal{E}$ by aligning their distribution $\mathcal{N}(\mu_\theta, \sigma^2_\theta)$ and $\mathcal{N}(\mu_\phi, \sigma^2_\phi)$.

Given the D set of training data in the form of $\{x_{t-1}, P_{t-1}, x_t, P_t\}_i^D$, our motion prior is trained by the following loss function composed of four terms:

$$L = \mathcal{L}_{rec} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{ct}\mathcal{L}_{ct} + \lambda_{SMPL}\mathcal{L}_{SMPL}. \tag{8}$$

with weights $\lambda_{KL} = 4e^{-4}$, $\lambda_{ct} = 0.01$, $\lambda_{SMPL} = 0.5$. The first term $\mathcal{L}_{rec}$ is the reconstruction loss that minimizes the difference between the decoder output and the ground truth:

$$\mathcal{L}_{rec} = \|\hat{x}_t - x_t\|^2 + 0.1 \times \|\hat{P}_t - P_t\|^2, \tag{9}$$

where $\hat{x}$, $\hat{P}$ are the predicted pose and phase features, respectively. The second term $\mathcal{L}_{KL}$ enforces the distribution learned by the learnable prior is close to the one from the encoder by measuring their KL-divergence. The third term $\mathcal{L}_{ct}$ is the contact loss

$$\mathcal{L}_{ct} = \sum_j BCE(\hat{c}_t^j, c_t^j) + \hat{c}_t^j \|\hat{v}_t^j\|^2 \tag{10}$$

where BCE is the binary cross-entropy; $\hat{c}_t^j$ and $\hat{v}_t^j$ represent the predicted contact label and the velocity of the j-th joint, respectively. This term encourages the decoder to predict correct contact labels while keeping the velocities of the joints in contact with the ground to become zero. The final term $\mathcal{L}_{SMPL}$ is optional which enforces the generated body meshes to be consistent with the SMPL model in the ground truth dataset, which is the same as one described in HuMoR [49].

4.4. Robust Inference of PhaseMP

Given an initial pose and phase features x_0 and P_0, an embedding for transition z_0 is sampled randomly from the prior distribution $\mathcal{R}(x_0, P_0)$, from which the decoder generates the change of pose and phase features, and then they are updated via Eq.6. This process can be performed repeatedly to generate a continuous motion $(x_0, x_1, x_2, \cdots)$. Although we observe that our phase-conditioned motion prior can generate better quality motions already when compared to other methods, the generated motions can still deteriorate especially when the input observation is highly unreliable due to noisy or missing joints. We thus propose a novel way to update the phase feature that considers the confidence of the input observation dynamically. It is computed as follows:

$$\begin{aligned} \bar{\boldsymbol{p}}_t &= \bar{\boldsymbol{A}}_t \cdot I(\alpha_P)(R(\theta) \cdot \boldsymbol{p}_{t-1}, (\boldsymbol{p}_{t-1} + \Delta \boldsymbol{p})) \\ \bar{\boldsymbol{A}}_t &= (1-\alpha_A)\boldsymbol{A}_{t-1} + \alpha_A(\boldsymbol{A}_{t-1} + \Delta \boldsymbol{A}) \\ \bar{\boldsymbol{F}}_t &= (1-\alpha_F)\boldsymbol{F}_{t-1} + \alpha_F(\boldsymbol{F}_{t-1} + \Delta \boldsymbol{F}) \\ \theta &= \Delta t \cdot 2\pi \cdot \bar{\boldsymbol{F}}_t \end{aligned} \tag{11}$$

where the update is basically performed by the interpolation of two sources, one from network prediction and the other from a cyclic update by rotating the phase manifold vector $\boldsymbol{p_{t-1}}$ with θ. $I(\cdot)$ refer to linear interpolation, $R(\cdot)$ is the rotation operation, Δt is a time-step between adjacent frames, and $\alpha_* \in [0, 1]$ represents how reliable each component of the phase feature is given the current pose feature. The confidence values α_A, α_P, and α_F are computed by:

$$\alpha_A, \alpha_P, \alpha_F = \begin{cases} 0, 0, 0 & \text{if} \quad \phi_t < \phi_{low} \\ \phi_t, \phi_t, \phi_t & \text{if} \quad \phi_{low} \leq \phi_t < \phi_{high} \\ 1, 0.5, 1 & \text{if} \quad \phi_t \geq \phi_{high} \end{cases} \tag{12}$$

where $\phi_t \in [0, 1]$ is the confidence value of the input observation at frame t (e.g., values given by 2D pose detectors), and ϕ_{low}, ϕ_{high} are user-specified minimum and maximum confidence thresholds where we use 0.4, 0.8, respectively. Note we set $\alpha_P = 0.5$ even when the input observation is fully reliable as demonstrated in [56]. If the confidence value is high ($\alpha_* = 1$) then the phase feature is updated following Eq. 11. If the confidence is low ($\alpha_* = 0$), the frequency and amplitude feature $\boldsymbol{A}_t, \boldsymbol{F}_t$ are carried over from the previous frame as the same as $\boldsymbol{A}_{t-1}$ $\boldsymbol{F}_{t-1}$, and are used to update the phase manifold features. Otherwise, the values are interpolated based on corresponding confidence values.

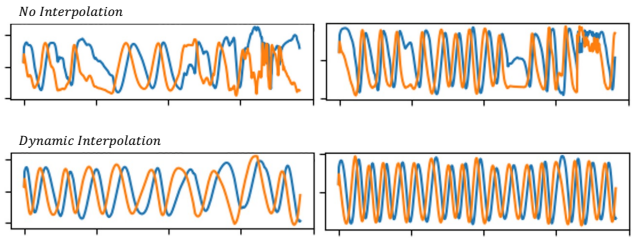

Figure 3. A visualization of the phase manifold vector $\boldsymbol{p}$ on two channels. The above represents the phase feature directly predicted by the network, while the bottom vector shows the interpolated phase based on Eq. 11. This shows that dynamic interpolation greatly helps to increase stability when the input pose features are noisy.

Implementation The prior and encoder networks are both designed as fully-connected 5-layer MLPs equipped with ReLU activation units and group normalization. The decoder is designed as a 4-layer SirenMLP [54] with sine activation of 60 sine factor. We follow the initialization scheme introduced in Siren [54]. Moreover, we introduce skip connections from the phase feature to each layer of the decoder network, further enhancing its influence. Following MotionVAE [38], we also use scheduled sampling to ensure the network learns from its own predictions. More details will be introduced in the supplemental materials.

5. Test-time Optimization

During run-time, given partial observations, such as 2D landmark sequence or partial 3D joint sequence, we estimate the original 3D pose sequence by optimization. In this section, we outline the details for integrating our phase-conditional motion prior with optimization.

There are three stages to perform the optimization. Given a sequence of observations $o_{1:T}$ (e.g., a sequence of 2D joint positions in a video) with length T, a ground plane is estimated and a sequence of SMPL poses are roughly fitted to the observations in Stage 1. An initial sequence of phase features $P_{1:T}$ is then computed from these poses in Stage 2 using the encoder in Eq. 7. In Stage 3, starting from the roughly fitted poses in Stage 1 and the initialized sequence of phase features in Stage 2, a refinement step is performed by using our PhaseMP to optimize the pose sequence with energies measuring plausibility of the pose sequence. All the above stages are performed by different optimization targets and energy terms.

We will describe Stage 3 here and explain the other stages in the supplementary materials. The optimization

problem in Stage 3 is formulated as follows:

$$\begin{aligned}\underset{z_{1:T-1},\beta,g}{\operatorname{argmin}} \; (&\lambda_{obs}E_{obs} + \lambda_{prior}E_{prior} \\ &+\lambda_{reg}E_{reg} + \lambda_{phase}E_{phase})\end{aligned} \tag{13}$$

where β is the SMPL shape parameter, and g is the ground plane parameter. The energy function comprises four terms where the first three terms are those proposed in HuMoR [49] while the last term is newly introduced in this work, which can significantly improve the motion quality especially for challenging scenarios. The weights can be set differently depending on the types of tasks, our settings for each task are included in Appendix. The details of each energy are explained below.

Observation Energy The purpose of the observation energy is to enforce the predicted sequence to align with the given observations:

$$E_{obs} = \sum_{t=1}^{T} ||\mathcal{O}(\hat{x}_t) - o_t||^2 \tag{14}$$

where o_t is an instance of observation at frame t and $\mathcal{O}$ is a function that projects the pose sequences $\hat{x}_{0\sim T}$ which are predicted by our model to the observation space. For instance, a 3D-to-2D projection can be used for the video-to-motion task while a masking function can be used for reconstructing full 3D poses from 3D partial markers.

Motion Prior Energy The goal of the motion prior term is to measure whether the given sequence of latent variables $z_{1:T}$ represents a plausible motion. This can be computed by

$$\begin{aligned}E_{prior} &= -\sum_{t=1}^{T-1} \log \mathcal{N}(\mathbf{z}_t; \mu_\theta^t, \sigma_\theta^t) \\ \mu_\theta^t, \sigma_\theta^t &= \mathcal{R}(x_t, P_t)\end{aligned} \tag{15}$$

where $\mu_\theta^t, \sigma_\theta^t$ are the mean and standard deviation of the prior distribution of pose transitions predicted by $\mathcal{R}$.

Regularization Energy The regularization energy helps the optimized motion to be smooth and consistent. It consists of four terms:

$$\begin{aligned}E_{reg} = \sum_{t=1}^{T} (&||\hat{J}_t - \hat{J}_t^{smpl}||^2 + ||\hat{l}_t - \hat{l}_{t-1}||^2 \\ &+\hat{c}_t^{foot}||v_t^{foot}||^2 + ||g - g_{init}||^2)\end{aligned} \tag{16}$$

where the first term regularizes the distance between the predicted joint positions $\hat{J}_t$ and the joint positions $\hat{J}_t^{smpl}$ computed from the SMPL pose parameters; the second term enforces the bone lengths l_t to be consistent over time; the third term makes the foot stationary by minimizing its velocity v^{foot} when the contact label $\hat{c}_t^{foot}$ is enabled; and the final term prevents the ground plane to deviate from its initial guess g_{init} during the optimization.

Phase-based Energy As demonstrated in previous work [49], natural-looking motions can be often generated when the input observation is dense and reliable, for example, when all joints are clearly visible in the video and the motion is not extremely dynamic. However, the quality of generated motions is significantly degraded when insufficient cues are provided, due to occlusions or the body being out-of-sight. Simply increasing the weights of the motion prior energy cannot handle invisibility for a long duration (>1s). Here, we introduce a novel phase-based energy to mitigate this challenge.

To calculate the energy, we first compute target phase features $\bar{P}_{1:T}$ by setting the initial phase to 0 and updating it by inference with Eq. 11. We then optimize the phase features $P_{1:T}$ by minimizing its difference with the target phase features $\bar{P}_{1:T}$:

$$E_{phase} = \sum_{t=1}^{T} ||\bar{P}_t - P_t||^2. \tag{17}$$

For the optimization, $P_{1:T}$ is first initialized by the phase features extracted from the SMPL poses in Stage 2. This energy can be considered as an additional regularization term in the frequency domain by encouraging the generated motion to maintain similar periodicity to ones computed from reliably observations only.

Implementation By default, we use L-BFGS optimization with a step size of 1 and a maximal number of iterations per optimization step of 20, which is implemented by PyTorch. It takes approximately 6 minutes to fit a 3-second sequence with a GTX 3090 graphics card. More details about optimization exist in the supplemental materials.

6. Evaluation

We evaluate our system on (i) motion reconstruction task, (ii) motion completion from sparse 3D joint markers, (iii) video-to-motion task in challenging scenes, (iv) ablation study. More qualitative and quantitative experiments are available in the supplemental materials.

6.1. Datasets and Metrics

Datasets The evaluation of our proposed method utilizes three human motion datasets. (1) **AMASS** [31], the largest dataset in terms of motion capture data, is curated from various sources and represented in the SMPL format, all the

Model	per-frame reconstruction				sampling (5s)			sampling (10s)		
	Contact ↑	MPJPE↓	PJPE-std↓	MV-PE↓	Contact ↑	ADE ↓	FDE ↓	Contact ↑	ADE ↓	FDE ↓
HuMoR [49](MLP, w/o Phase)	0.9770	0.022	0.051	0.057	0.8585	36.14	47.34	0.8216	45.43	62.47
Ours(MLP, with Phase)	0.9764	0.020	0.040	0.061	0.8646	32.36	35.73	0.8525	39.48	54.96
Ours(SirenMLP, w/o Phase)	0.9788	0.019	0.031	**0.044**	0.8691	**31.81**	**35.59**	0.8577	**35.47**	49.95
Ours(SirenMLP, with Phase)	**0.9799**	**0.017**	**0.021**	0.047	**0.8702**	34.88	36.91	**0.8662**	42.12	**49.47**

Table 1. Comparison of different methods on the per-frame prediction (Left) and random sampling with different durations (Middle, Right). The MPJPE and DE are measured as positional errors with the unit of centimeters. For the sampling experiments, we use the same initial pose from ground truth, and then run the sampling 50 times, and choose the one with the lowest ADE (average) to report its FDE(final frame).

data are down-sampled to 30 fps in our experiments. We train the phase extractor using the CMU subset and then train the phase-conditioned variational autoencoder (VAE) on the training subsets which are the same as [49, 61]. All evaluation is conducted using the Transitions and HumanEva subsets. (2) **i3DB** [43] and (3) **PROX** [16], which contain RGB videos of human-environment interaction and are used for a comprehensive quantitative and qualitative evaluation of the video-to-motion task. We first perform pre-processing by running Openpose, human mask detection, and plane estimation to facilitate further test-time pose estimation.

Baseline and Metrics Based on our inspiration from the HuMoR baseline, we propose several enhancements in our system, including a conditional feature, a new network module, and different optimization energy terms. We evaluate the effectiveness of these improvements in challenging scenarios through experiments.

To assess the accuracy of our reconstruction, we compute the commonly used metrics MPJ-PE and MV-PE [42, 31, 49], to show the mean positional distance (cm) of body joints and vertices between our prediction and the ground truth. Additionally, we use the displacement distance (DE) between the generated motion and ground truth, to evaluate whether our learned prior can effectively reconstruct the desired motion. To examine the physical realism of the motion estimation task [38, 49], we also measure the accuracy of contact (Contact), mean per-joint accelerations (Accel), foot-ground penetrations with a 15cm threshold, penetration occurrence frequency (P-Frep), mean penetration distance (P-Dis).

6.2. Evaluation of Motion Reconstruction

Because our model is based on conditional VAEs, we first evaluate our model by measuring the accuracy on reconstruction (prediction). The accuracy is evaluated by two criteria, where we use the AMASS dataset to train and test our model. The first criterion is the per-frame prediction accuracy, where the ground truth input of x_{t-1}, P_{t-1}, and the latent variable z_t obtained from $E(x_{t-1}, x_t)$ are given. The second criterion evaluates the sequential output by initializing the system with x_0 and P_0 from a test motion, and then sampling 50 different motion sequences autoregressively from the same seed for the same length as the test motion. We select the sequence with the lowest average displacement error (ADE) as the closest prediction. We report both the ADE and last frame distance (FDE) for the selected sequence. The results are presented in Table 1 which shows that our full method (SirenMLP+Phase) achieves the best performance on average and all other variants also outperform the SOTA baseline.

6.3. Estimation from 3D Observations

We conduct an experiment to evaluate the effectiveness of test-time optimization in filling incomplete 3D joints. To simulate real-world occlusions, we generate three types of inputs: 1) occluded joints in time (missing frames) or in space (joints above 0.9m in height are only visible; or the end-effectors are only visible); 2) joint positions with noise. We then recover the original pose sequence using our *PhaseMP*. The performance is evaluated based on the average positional error for visible (**Vis**) and occluded (**Occ**) joints.

The experimental results are presented in Table 2 and Table 3. Our approach outperforms other methods in predicting occluded joints, as demonstrated by the average positional error of all and invisible (**Occ**) joints. Furthermore, our method can predict contact labels more accurately than HuMoR. Regarding the denoising experiments, our approach produces smoother results, particularly when the degree of noise is high.However, an interesting observation is that our method does not align visible joints as accurately as VPoser-t. This can be attributed to the additional constraints imposed by the periodic manifold, which encourages more realistic motion rather than focusing solely on local alignments.

Some visualization comparison is shown in Figure 5 where our method generates more natural-looking and dynamic motions with less foot sliding when compared to HuMoR. The differences are best seen in the supplemental video.

Method	Input Conditions	fitting (3s)							fitting (5s)						
		Vis	Occ	All	Contact ↑	Accel	P-Frep	P-Dis	Vis	Occ	All	Contact↑	Accel	P-Frep	P-Dis
VPoser-t	J_{height}>0.9	**0.67**	20.76	9.22	-	5.71	16.77%	2.28	-	-	-	-	-	-	-
MVAE		2.39	19.15	9.52	-	7.12	3.15%	0.30	-	-	-	-	-	-	-
HuMoR		1.46	17.40	8.24	**0.89**	5.38	3.31%	**0.26**	**2.38**	18.44	9.68	0.85	4.87	4.92%	**0.30**
Ours		3.94	**15.63**	**8.31**	**0.89**	**4.58**	**3.04%**	0.28	3.29	**16.08**	**8.41**	**0.87**	**4.69**	**4.32%**	0.31
HuMoR	$J_{\text{end\_effectors}}$	**3.05**	4.12	3.83	0.96	4.91	0.31%	1.03	**3.15**	4.20	3.91	0.96	**4.97**	1.19%	1.25
Ours		3.16	**4.07**	**3.79**	**0.97**	**4.88**	**0.28%**	**1.02**	3.32	**4.16**	**3.88**	0.96	4.99	**0.82%**	**1.13**
HuMoR	10 frames interval	5.56	7.76	7.49	0.91	7.72	1.57%	1.90	11.04	13.53	13.25	0.87	9.65	19.02%	5.97
Ours		**3.19**	**4.92**	**4.33**	**0.93**	**6.33**	**1.31%**	**1.72**	**8.48**	**10.21**	**9.30**	**0.92**	**8.42**	**9.73%**	**3.73**

Table 2. Comparison of different methods on different input settings: 1) Incomplete joints; 2) Root and 5 end-effectors; 3) Keyframes with 10 frame intervals. Comparisons are performed with two different durations (3 and 5 seconds). All the results are measured by two groups of metrics, where one group is the positional error for visible/occluded/all joints, and the other group is used to measure the physical realism including the contact accuracy and foot-ground penetrations statistics with 15cm threshold.

Method	Noisy Radius	All	Contact↑	P-Frep	P-Dis
1D-Filter	4cm	3.91	-	2.45 %	0.14
VPoser-t	4cm	3.67	-	1.35%	0.07
MVAE	4cm	2.68	-	1.75%	0.11
HuMoR	4cm	2.27	0.97	1.18%	**0.05**
Ours(w/o Phase)	4cm	2.12	0.97	1.22%	0.06
Ours(with Phase)	4cm	**1.96**	**0.98**	**1.14%**	**0.05**
1D-Filter	12cm	11.89	-	4.87%	2.66
HuMoR	12cm	34.08	0.77	7.29%	5.26
Ours	12cm	**9.42**	**0.89**	**4.20%**	**0.48**

Table 3. Motion estimation from noisy inputs, where two different noise levels are tested.

6.4. Estimation from RGB Observations

We evaluate the proposed method in the video-to-motion task, where the goal is to estimate 3D pose sequences from RGB videos. We use i3DB [43] and PROX [16] datasets, and also obtain the 2D poses and confidence values from OpenPose [7]. To mitigate the local minima issue, we follow the same procedure proposed by HuMoR, which splits the entire motion into 3-second sub-sequences. However, instead of stitching all sub-sequences after the parallel optimization, we optimize them one by one, using the last frame in the optimized sequence s_{t-1} as the initial pose of s_t. As a result, we can easily stitch them sequentially to obtain a full-frame reconstruction. The results are presented in Table 4. It is observed that optimization-based methods produce more accurate results, and our proposed method further improves the estimation compared to others. We also show qualitative results in Figure 4. Our method maintains consistency in long sequences thanks to the temporal features captured by the phase feature, which stabilizes the estimation in heavily occluded frames, such as sitting on a sofa. More results for the video-to-motion task can be found in our supplementary materials.

Method	i3DB				PROX	
	MPJPE	P-MPJPE	P-Frep	P-Dis	P-Frep	P-Dis
VIBE	116.46	15.08	7.98%	3.01	23.46%	4.71
VPoser-t	32.73	16.62	9.59%	2.68	13.38%	2.82
MVAE	40.91	19.17	7.43%	1.55	-	-
HuMoR	28.15	14.51	2.12%	0.68	9.99%	1.56
Ours	**27.43**	**14.19**	**2.03%**	**0.61**	**9.12%**	**1.38**

Table 4. Comparison of different methods for the video-to-motion task on i3DB [43] and PROX [16] datasets. The P-MPJPE is used for calculating the positional error after root alignment.

6.5. Ablation Study

Here we analyze the effectiveness of each newly-added component of our system for the motion reconstruction task. The optimization without any new component is used as the baseline. Then we compare the results of the reconstructed motion w/wo SirenMLP (SI), w/wo Phase Condition (PC), w/wo Dynamic Test-Time Interpolation(DI). The experiment is done with i3DB dataset in the same setting described in Sec. 6.4. The results are shown in Tab. 5.

Method	Components			MPJPE	P-MPJPE	P-Frep	P-Dis
VIBE	-			116.46	15.08	7.98%	3.01
Baseline	-			30.49	15.44	2.61%	0.79
Ours	SI ×	PC ✓	DI ×	31.75	18.33	2.51%	0.83
Ours	SI ✓	PC ×	DI ×	29.91	15.42	2.73%	0.74
Ours	SI ×	PC ✓	DI ✓	27.82	14.90	2.39%	0.64
Ours	SI ✓	PC ✓	DI ✓	**27.43**	**14.19**	**2.03%**	**0.61**

Table 5. Comparison of different system components involved setting.

The ablation study reveals the effectiveness of different components in our framework. Phase is a powerful signal which helps to reproduce realistic movements. However, using it solely will result in an unstable prediction under heavy occlusion scenarios due to the error accumulation through the auto-regressive phase update. Dynamic interpolation masks out the low confidence frames, and its combination with the phase makes the prediction more robust in challenging cases. The SirenMLP module also improves the baseline results.

Figure 4. Qualitative comparison for video-to-motion task. We run video-based motion reconstruction method VIBE [31], test-time optimization with HuMoR [49] and our method in the same settings and report the results. All the methods are trained with AMASS [31] dataset only. Our method can produce more coherent and realistic motion even in the existence of heavy occlusion.

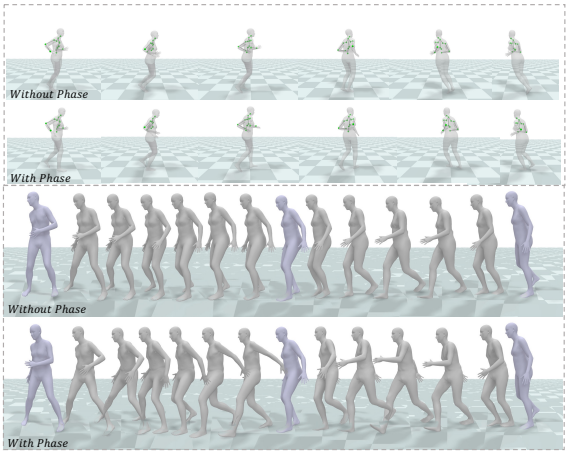

Figure 5. Motion generation from incomplete input. The first and second rows are the fully body reconstruction from the upper body only with/without using the phase feature. The bottom two rows show motion-in-betweening results from sparse keyframes shown as purple color.

7. Discussion

We have demonstrated the benefits of our phase-conditioned motion prior model, **PhaseMP**, in a variety of 3D pose estimation tasks under challenging settings. It is a general motion representation model that can be applied to encode not only periodic actions such as locomotion but also complex non-periodic actions such as dancing.

Our system has several limitations. Firstly, it relies on an assumption the motions are performed on flat ground and the cameras are static. Secondly, our system might fail when the occlusion period is excessively long because the ambiguity in the phase prediction increases. Thirdly, though most movements can be reconstructed with the periodic autoencoder, the diversity of the output motions may be restricted to those observed in the training data. Lastly, as same with other methods, our optimization process can still encounter failures when tested on motions that significantly deviate from the training data.

One intriguing future direction could involve utilizing the learned phase feature as a positional encoding for a transformer architecture, allowing the phase feature to directly influence predictions without test-time optimization. Another promising avenue for future research would be exploring the integration of multi-modal signals that can be easily combined in the frequency domain, such as sound waves from input videos.

Acknowledgments

This work was mainly done during Mingyi Shi's internship in Meta. We thank Deepak Gopinath for his valuable input. Jungdam Won was partially supported by the New Faculty Startup Fund from Seoul National University, ICT(Institute of Computer Technology) at Seoul National University, and the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2020-0-01460) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). Taku Komura and Mingyi Shi are partly supported by Technology Commission (Ref:ITS/319/21FP) and Research Grant Council (Ref: 17210222), Hong Kong.

References

[1] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceeding of the IEEE/CVF Conference in Computer Vision and Pattern Recognition*, CVPR'19, pages 3395–3404, June 2019.

[2] Andreas Baak, Meinard Müller, Gaurav Bharaj, Hans-Peter Seidel, and Christian Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *2011 International Conference on Computer Vision*, pages 1092–1099, 2011.

[3] Philippe Beaudoin, Pierre Poulin, and Michiel van de Panne. Adapting wavelet compression to human motion capture clips. In *Proceedings of Graphics Interface 2007*, pages 313–318, 2007.

[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 *Proceedings of the European Conference on Computer Vision*, ECCV '16, page 561–578, Berlin, Germany, 2016. Springer.

[5] Armin Bruderlin and Lance Williams. Motion signal processing. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '95, page 97–104, New York, NY, USA, 1995. Association for Computing Machinery.

[6] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2272–2281, 2019.

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '18, Washington, DC, USA, 2018. IEEE Computer Society.

[8] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation in videos. *arXiv preprint arXiv:2002.10322*, 2020.

[9] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T. Tan. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[10] H. Ci, X. Ma, C. Wang, and Y. Wang. Locally Connected Network for Monocular 3D Human Pose Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[11] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *CVPR*, 2023.

[12] A. Elgammal and Chan-Su Lee. Separating style and content on a nonlinear manifold. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I, 2004.

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

[14] S. Ghorbani, C. Wloka, A. Etemad, M. A. Brubaker, and N. F. Troje. Probabilistic character motion synthesis using a hierarchical deep latent variable model. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '20, Goslar, DEU, 2020. Eurographics Association.

[15] Julian Habekost, Takaaki Shiratori, Yuting Ye, and Taku Komura. Learning 3d global human motion estimation from unpaired, disjoint datasets. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

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

[17] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7779–7788, 2020.

[18] M. Heideman, D. Johnson, and C. Burrus. Gauss and the history of the fast fourier transform. *IEEE ASSP Magazine*, 1(4):14–21, 1984.

[19] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics*, 39(4):236:1–236:14, 2020.

[20] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *European Conference on Computer Vision*, pages 69–86. Springer, 2018.

[21] Nicholas R. Howe, Michael E. Leventon, and William T. Freeman. Bayesian reconstruction of 3d human motion from single-camera video. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 820–826, Cambridge, MA, USA, 1999. MIT Press.

[22] Wenbo Hu, Changgong Zhang, Fangneng Zhan, Lei Zhang, and Tien-Tsin Wong. Conditional Directed Graph Convolution for 3D Human Pose Estimation. *arXiv:2107.07797 [cs]*, 2021.

[23] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, July 2014.

[24] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable Triangulation of Human Pose. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7717–7726, Seoul, Korea (South), 2019. IEEE.

[25] 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*, CVPR '18, pages 7122–7131, Washington, DC, USA, 2018. IEEE Computer Society.

[26] Angjoo Kanazawa, Jason Y. JZhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR'19, 2019.

[27] Ben Kenwright. Quaternion fourier transform for character motions. In *Workshop on Virtual Reality Interactions and Physical Simulations*, 2015.

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

[29] Diederik P Kingma and Max Welling. Learning structured output representation using deep conditional generative models. In *International Conference on Machine Learning*, pages 348–356. PMLR, 2014.

[30] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[31] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262. IEEE, June 2020.

[32] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Muller, Otmar Hilliges, and Michael J Black. SPEC: Seeing People in the Wild With an Estimated Camera. page 11.

[33] 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 International Conference on Computer Vision*, 2019.

[34] 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 International Conference on Computer Vision*, ICCV'19, 2019.

[35] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[36] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021.

[37] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 2022.

[38] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. Character controllers using motion vaes. *ACM Trans. Graph.*, 39(4), 2020.

[39] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[40] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, Oct. 2015.

[41] Haoyu Ma, Zhe Wang, Yifei Chen, Deying Kong, Liangjian Chen, Xingwei Liu, Xiangyi Yan, Hao Tang, and Xiaohui Xie. Ppt: token-pruned pose transformer for monocular and multi-view human pose estimation. In *ECCV*, 2022.

[42] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceeding of the IEEE International Conference on Computer Vision*, ICCV '17, pages 2659–2668, Oct 2017.

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

[44] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recogßnition (CVPR)*, June 2021.

[45] Dirk Ormoneit, Hedvig Sidenbladh, Michael Black, and Trevor Hastie. Learning and tracking cyclic human motion. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

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

[47] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '19, Washington, DC, USA, 2019. IEEE Computer Society.

[48] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. In *SIGGRAPH Asia 2018 Technical Papers*, page 178. ACM, 2018.

[49] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021.

[50] Chris Rockwell and David F. Fouhey. Full-body awareness from partial observations. In *ECCV*, 2020.

[51] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *Computer Vision–ECCV 2022: 17th European Con-*

ference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V, pages 461–478. Springer, 2022.

[52] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. MotioNet: 3D Human Motion Reconstruction from Monocular Video with Skeleton Consistency. *ACM Transactions on Graphics*, 40(1):1:1–1:15, 2020.

[53] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Trans. Graph.*, 40(4), jul 2021.

[54] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*, 2020.

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

[56] Sebastian Starke, Ian Mason, and Taku Komura. DeepPhase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)*, 41(4), 2022.

[57] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *IEEE International Conference on Computer Vision, ICCV*, 2021.

[58] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021.

[59] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022.

[60] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *IEEE International Conference on Computer Vision, ICCV*, 2019.

[61] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, October 2022.

[62] Munetoshi Unuma, Ken Anjyo, and Ryozo Takeuchi. Fourier principles for emotion-based human figure animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 91–96, 1995.

[63] Raquel Urtasun, David J. Fleet, and Pascal Fua. Temporal motion models for monocular and multiview 3d human body tracking. *Computer Vision and Image Understanding*, 104(2):157–177, 2006. Special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour.

[64] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.

[65] Zhe Wang, Daeyun Shin, and Charless C. Fowlkes. Predicting Camera Viewpoint Improves Cross-dataset Generalization for 3D Human Pose Estimation. *arXiv:2004.03143 [cs]*, 2020.

[66] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, June 2022.

[67] Heyuan Yao, Zhenhua Song, Baoquan Chen, and Libin Liu. Controlvae: Model-based learning of generative controllers for physics-based characters. *ACM Trans. Graph.*, 41(6), nov 2022.

[68] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. *arXiv preprint arXiv:2212.02500*, 2022.

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

[70] M Ersin Yumer and Niloy J Mitra. Spectral style transfer for human motion between independent actions. *ACM Transactions on Graphics (TOG)*, 35(4):1–8, 2016.

[71] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13232–13242, June 2022.

[72] Siwei Zhang, Yan Zhang, Federica Bogo, Pollefeys Marc, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *International Conference on Computer Vision (ICCV)*, Oct. 2021.

[73] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision (3DV)*, Nov. 2020.

[74] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, , and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European conference on computer vision (ECCV)*, Oct. 2022.

[75] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic Graph Convolutional Networks for 3D Human Pose Regression. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3420–3430, 2019.

[76] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[77] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3D Human Pose Estimation With Spatial and Temporal Transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, page 10, 2021.

[78] Zhiming Zou and Wei Tang. Modulated graph convolutional network for 3d human pose estimation. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision (ICCV), pages 11477–11487, October 2021.