H4D: Human 4D Modeling by Learning Neural Compositional Representation

Boyan Jiang[1*] Yinda Zhang[2*] Xingkui Wei[1] Xiangyang Xue[1] Yanwei Fu[1]
[1]Fudan University [2]Google

Abstract

Despite the impressive results achieved by deep learning based 3D reconstruction, the techniques of directly learning to model 4D human captures with detailed geometry have been less studied. This work presents a novel framework that can effectively learn a compact and compositional representation for dynamic human by exploiting the human body prior from the widely used SMPL parametric model. Particularly, our representation, named H4D, represents a dynamic 3D human over a temporal span with the SMPL parameters of shape and initial pose, and latent codes encoding motion and auxiliary information. A simple yet effective linear motion model is proposed to provide a rough and regularized motion estimation, followed by per-frame compensation for pose and geometry details with the residual encoded in the auxiliary code. Technically, we introduce novel GRU-based architectures to facilitate learning and improve the representation capability. Extensive experiments demonstrate our method is not only efficacy in recovering dynamic human with accurate motion and detailed geometry, but also amenable to various 4D human related tasks, including motion retargeting, motion completion and future prediction.

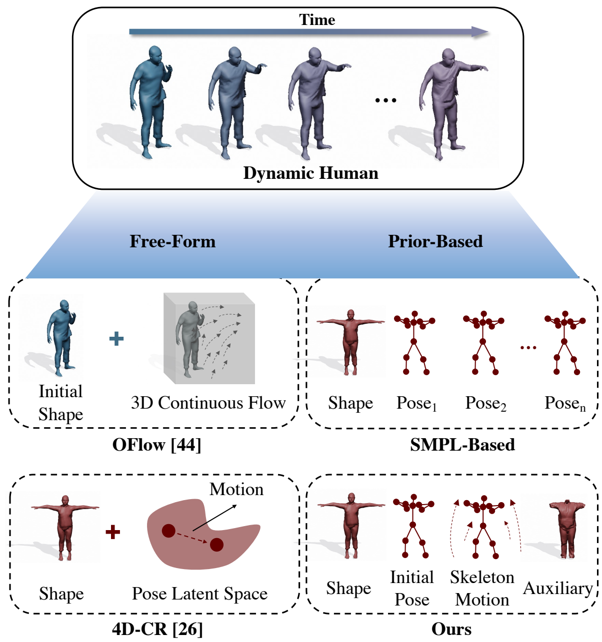


Figure 1. **Comparison with existing 4D human representations.** Our representation supports faster inference and more complete reconstructions compared with free-form methods (Fig. 3). And it provides long-range temporal context and additional fine-grained geometry controlled by low-dimensional SMPL parameters and latent codes, which is more compact compared with previous SMPL-based methods.

1. Introduction

The vanilla SMPL based parametric representations have been extensively studied and widely utilized for modeling 3D human shapes, and thus shown critical impacts to many human-centric tasks, such as pose estimation [16, 24, 30, 32, 34, 42] and body shape fitting [9, 18, 33, 48, 58]. However, these representations are arguably insufficient for applications involving dynamic signals, *e.g.* 3D moving humans (Fig.1 top), since the temporal information is not captured.

As solutions, 4D representations are proposed and can be in general categorized into free-form and prior-based methods depending on the 3D representation of the output shape (Fig. 1). The free-form methods leveraging Neural ODE [13] and deep implicit function [26, 44] often rely on computational expensive architectures to learn the compact latent spaces and reconstruct 4D sequences. Unfortunately, since the human body prior is not explicitly modeled, the reconstruction results of these methods may contain obvious geometry artifacts such as missing hands, and their modeling errors accumulate rapidly over time. On the other hand, prior-based methods [30, 32, 66] are mostly derived from the SMPL parametric model [37], which typically employs one shape parameter and a series of pose parameters to model dynamic sequences. Although they produce plausible results, their motion representations are not compact or only support a small time span, *e.g.* $\pm$ 5 frames [30].

* indicates equal contributions.
Boyan Jiang, Xingkui Wei and Xiangyang Xue are with the School of Computer Science, Fudan University.
Yanwei Fu is with the School of Data Science, Fudan University, and Fudan ISTBI—ZJNU Algorithm Centre for Brain-inspired Intelligence, Zhejiang Normal University, Jinhua, China.

In this paper, we propose **H4D**, which is a novel neural representation for **H**uman **4D** modeling that combines the merits of both the prior-based and free-form solutions. To reflect the compositional natures [60], we encode each dynamic human sequence with SMPL parameters representing shape and initial pose, and a compact latent code representing temporal motion, which can then be used to reconstruct the input sequence through a decoder. At the core of this decoder, a simple yet effective prior model extended from SMPL [37] is designed to provide coarse but long-term estimation of the 3D human geometry and motion. This can ensure more complete and plausible outputs compared to the prior arts of free-form reconstructions [26, 44], but potentially be inclined to suffer from the limited representation capability. To this end, we add an additional auxiliary latent code to our representation to compensate the inaccurate motion and enrich the geometry details. Such a representation takes full advantage of parametric models by exploiting strong prior based regularization for plausible initialization and complement them with powerful deep learning components to facilitate the human 4D modeling with impressive motion and geometry accuracy.

Our representation is learned via an auto-encoding framework. The encoder predicts the SMPL parameters and latent codes for each aspect from densely sampled point clouds, which are fed into the decoder to reconstruct the identical input dynamic human sequence. Once trained, the encoder and decoder are both fixed to support various applications, such as motion retargeting, completion, and prediction, through either forward propagation (feed-forward) or backward optimization (auto-decoding) depending on the inputs. We design novel Gated Recurrent Unit (GRU) [15] based architectures for both encoder and decoder to benefit the model performance while working in either mode. In feed-forward mode, we do not require the input point clouds to be temporarily tracked, *i.e*. the point trajectories like in previous work [26, 44]. This simplifies the training requirements and enhances the applicability of high-level applications. In auto-decoding mode, our model leverages the temporal information for optimization, which is critical for robustness to recover detailed motion and geometry.

Contributions We propose H4D, a compact and compositional representation for 4D human captures, which combines a linear prior model with the residual encoded in a learned auxiliary code. The framework is learned via 4D reconstruction, and the latent representation can be extracted from either nonregistered point clouds in a feed-forward fashion or auto-decoding through optimization. Extensive experiments show that our representation and GRU-based architecture are effective in recovering accurate dynamic human sequences and providing robust performance for a variety of 4D human related applications, including motion retargeting/completion and future prediction.

2. Related Work

4D Representation There has been a lot of work aiming to reconstruct 3D objects based on various representations, such as 3D voxels [17, 22, 63], point clouds [1, 20, 50, 51], meshes [12, 23, 29, 36, 62] and implicit functions [11, 14, 19, 27, 43, 46]. However, the deep representation for 4D data, *i.e.* a time-varying 3D object, has received less attention mostly due to the challenge of encoding the temporal dimension. Pioneer work mostly relies on Neural ODE [13] and combines with occupancy network [43, 44], point clouds [53], and compositional property [26]. Despite state-of-the-art performance in various motion-related tasks, Neural ODE tends to accumulate errors over time that causes incomplete geometry, and slows down training convergence and inference run-time. In contrast, our model relies on the prior model for comprehensive geometry and motion and the recurrent network for efficient inference.

Human Body Estimation For human shape and pose estimation [9, 16, 24, 32–34, 42, 48] or motion prediction [2, 3, 7, 10, 39, 40], most of works are based on SMPL or its extension [37, 47, 54]. Specifically, HMMR [30] learns to encode temporal information by reconstructing a small number of past and future frames. Zhang *et al*. [66] propose the first autoregressive model for predicting 3D human motion from image sequences. VIBE [32] leverages GRU to regress SMPL parameters, and designs an adversarial learning framework to predict temporal transitions. Though produce plausible motion, the motion representations in these methods are either implicit [32], coupled with geometry [30, 32, 66], or limited to a short temporal range [30]. In contrast, we formulate the motion with a prior model based on PCA [28, 61] (its non-linear extension PGA [21] is also applicable) for long-range context followed by per-frame adjustments controlled by a learned latent code, which is compact, compositional, and tolerant to error accumulation.

Fine-grained Human Reconstruction Many human reconstruction methods [9, 29, 30, 32, 33, 42] are limited to the unclothed body as SMPL-based models may suffer from limited expressive power in the shape space. To capture fine-grained geometry, such as clothing or hair, the neural implicit function has been used to reconstruct a free-form surface [8, 14, 18, 55–57, 64, 67], but it is still challenging to recover detailed structure like fingers, face, or wrinkles on clothes. Another family of approaches extends the parametric model by predicting per-vertex displacements upon the canonical body mesh [4–6, 35, 38, 65], which achieves a good balance between the expressiveness and prior regularization. Most related to us, CAPE [38] trains a generator to synthesize fine-grained geometry from a latent space, and can run in the auto-decoding mode for fitting. However, it is empirically not robust to work with temporal frames and sensitive to the errors in imperfect poses, which is not ideal to plug and play in our scenario.

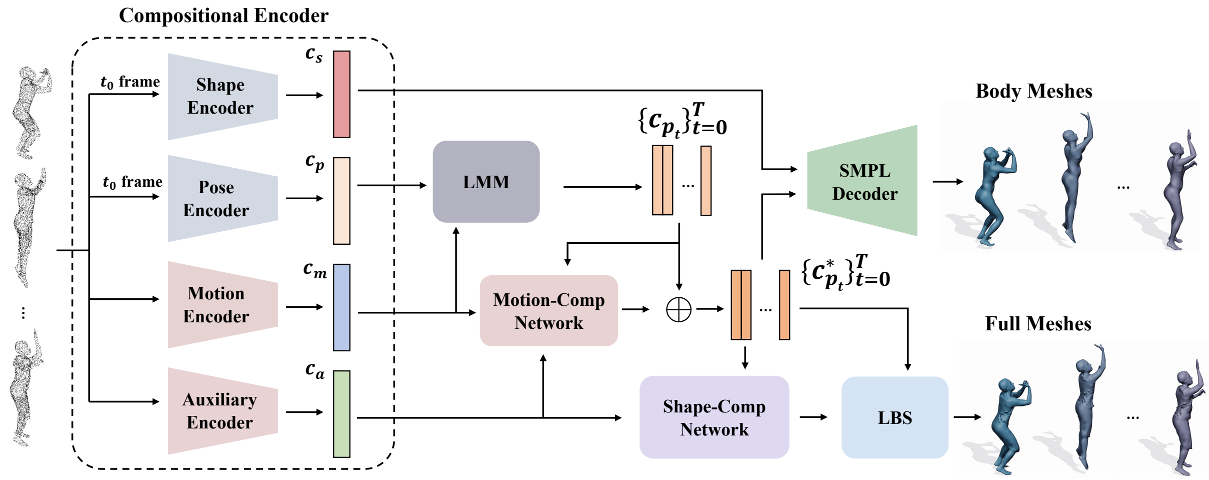


Figure 2. **Overview of our framework.** We learn the compositional representation for dynamic humans through 4D reconstruction. Specifically, given an input point cloud sequence, we first extract SMPL parameters of shape and initial pose, and latent codes of motion and auxiliary information with the compositional encoder, and obtain a rough motion estimation with the Linear Motion Model (LMM). Then we predict the residuals of temporal motion and shape in canonical pose with the GRU-based Motion-Comp Network and Shape-Comp Network, respectively. Our method is able to output accurate body mesh sequences with the SMPL decoder [37] as well as full mesh sequence with cloth and hair by using Linear Blend Skinning (LBS). Detailed architectures are in Supp. Mat.

3. Method

This section introduces our H4D representation, which is learned through reconstruction task (Fig. 2). Given a 3D human model performing motion in a time span (mesh sequence of 30 frames), we sample a point cloud of 8192 points from each as the input sequence to the network. Note that we do not assume the temporal correspondences among frames (*e.g.* point trajectories) are available, which are critically required by previous 4D representations [26, 44].

The input sequence is fed into a compositional encoder to extract SMPL parameters representing shape and initial body pose, together with latent codes representing temporal motion and auxiliary of additional compensation on motion and geometry (Sec. 3.1). To reconstruct the input temporal sequence, the shape, initial pose, and motion codes are combined through a pre-learned Linear Motion Model (LMM) to generate a rough estimation of per-frame 3D shapes represented as SMPL [37] (Sec. 3.2). Due to the limited capacity of LMM, the output, though plausible, demands additional refinements for accurate reconstruction. To this end, we feed the motion code, auxiliary code, and initial estimation to the GRU based Motion-Comp network (Sec. 3.3) and Shape-Comp network (Sec. 3.4) to predict the residual on temporal motion and shape in the canonical pose, respectively. The final sequence is obtained by deforming the refined canonical shape using the linear blending weight according to the refined per-frame poses.

3.1. Compositional Encoder

To keep the representation compositional, we train four separate encoders to extract SMPL parameters representing the shape c_s and initial pose c_p, and latent codes representing motion c_m and auxiliary information c_a. The shape and pose encoders are implemented as PointNet-based [51] network with ResNet blocks, which take only the starting frame as input since it is sufficient to tell the canonical body shape and initial pose. On the other hand, the motion and auxiliary encoders take all frames as input since temporal information is needed. To achieve that, we firstly encode the point cloud of each frame into a feature vector using a shallow PointNet, and then further aggregate per-frame feature with a GRU layer. The feature extractor is shared between motion and auxiliary encoders, and only GRUs are trained respectively. Note that our temporal encoders can process sequences of unordered point clouds without temporal correspondences.

3.2. Linear Motion Model

We take the predicted c_p and c_m to reconstruct a coarse estimation of motion. Inspired by Urtasun *et al.* [61], we employ the parameter space of SMPL model and pre-learn a linear model for the motion to ensure robustness. Each input temporal sequence can be represented as $\Phi = [\theta_1, \dots \theta_L], L = 30$, where $\theta_i \in \mathbb{R}^{72}$ is the SMPL pose parameter for frame i. We then represent motion as the per-frame difference of the pose parameter from the first frame, i.e. $\Psi = [\theta_2 - \theta_1, \dots, \theta_L - \theta_1] \in \mathbb{R}^{72(L-1)}$, and run a Principal Component Analysis (PCA) [28] to reduce the dimension. The input motion now can be reconstructed through the linear model: $\hat{\Phi} = [\theta_1, \alpha^T \cdot \mathcal{M} + \mu_\Psi + \theta_1]$, where $\alpha \in \mathbb{R}^K$ is the coefficient of principal components, $\mathcal{M} = [\mathbf{M}_1, \dots, \mathbf{M}_K] \in \mathbb{R}^{72(L-1)\times K}$ and μ_Ψ are the top K principal components and mean of Ψ.

In practice, we found it more robust to run PCA separately for the global orientation (*i.e.* pelvis) and body joint rotation. We pick 4 bases for global orientation and 86 basis for body joint rotation, which explains 90% of the variance[1]. Finally, we plug the linear motion model into our pipeline, amenable to the output of the compositional encoder: $\{c_{p_t}\}_{t=0}^{L-1} = [c_p, c_m^T \cdot \mathcal{M} + \mu_\Psi + c_p]$, where c_{p_t} is the pose parameter for frame t.

3.3. Motion Compensation Network

The LMM is effective in representing motion with a relatively large number of temporal frames; unfortunately, it lacks the capacity to represent motion details. As a result, the predicted pose sequences are not accurate enough.

To improve the motion accuracy, we build a motion compensation network (Motion-Comp) to adjust the pose parameter of each frame. Specifically, we adopt a GRU-based network [15] as it was demonstrated to be effective for processing temporal information. We concatenate the motion code c_m and the auxiliary code c_a to the pose parameters of each frame from LMM prediction $\{c_{p_t}\}_{t=0}^{L-1}$, and then feed them sequentially into GRU to produce the residual of each frame. Once the per-frame pose parameters are updated with the output of Motion-Comp network, we combine them with the shape parameter c_s from the encoder into the standard SMPL decoder to reconstruct the per-frame mesh. Overall, our motion model benefits from both the strong prior in the linear motion model and the impressive capacity of the motion compensation network.

3.4. Shape Compensation Network

So far, we are able to reconstruct the correct motion sequences, which can be further converted to body mesh sequences via SMPL decoder. However, the predicted shapes are still inferior, as many details such as hairs or clothes are missing. This is mostly due to the constrained capacity of SMPL shape space. To enhance the geometry, the shape representation presented in CAPE [38] is introduced: a per-vertex offset is estimated for the body mesh in the canonical space via a graph-based neural network conditioned on target pose. The added details would be then transferred to the target body pose via the pre-defined linear blending weight in SMPL. When combined with our framework, one straightforward way is to have the auxiliary code c_a encode shape details and feed it through the CAPE decoder for per-vertex offsets. We found this works reasonably well in feed-forward mode but not the back-propagation. We suspect this might because of the inconsistent gradients from different temporal frames, especially when the pose estimation is not perfectly accurate. As a result, the compensated geometry is vaguely correct (*e.g.* bump on the head for some hairstyles) but not precise. To improve the stability, we propose a shape compensation network (Shape-Comp), in which a GRU takes the auxiliary code c_a as input and predicts a new latent vector for each temporal frame conditioned on the predicted pose. The latent vector is then fed into the graph network to predict per-vertex offset, which is similar to the CAPE decoder. We remove the VAE and adversarial loss as they empirically hurt the performance. The GRU enables information exchanging across temporal frames, which is critical for robust back-propagation when running applications like motion completion and prediction.

[1]Please refer to Supp. Mat. for visualization of principal components

3.5. Training Strategy

Neural networks with multiple stages are highly nonlinear and could easily fall in the local minimum. We advocate a stage-wise training strategy to enhance the training stability. Specifically, we first train the shape encoder, pose encoder, point feature extractor and motion encoder jointly with the pre-learned linear motion model. Once the model converges, we enable the Motion-Comp and Shape-Comp networks for the end-to-end joint training. Similar training strategy has been commonly used by other works, *e.g.* BC-Net [25], XNect [41] and Predicting Human Dynamics [66].

Loss Functions Since our reconstructions are registered with SMPL topology, we use per-vertex L1-loss with the ground truth mesh as the objective function. To further alleviate the ambiguity between body shape and clothing, we add an L2-loss on the shape code c_s with regard to the ground truth. During the first training stage, we use the mesh reconstructed with the motion from LMM for supervision. In the second stage, we use added loss on both meshes before and after the Shape-Comp network. The detailed formulation of the loss functions can be found in Supp. Mat.

4. Experiments

In this section, we perform extensive experiments to verify the efficacy of our method. First, we evaluate the capacity of our representation for encoding accurate shape and motion on the tasks of 4D reconstruction and human shape and motion recovery. We then demonstrate that a large variety of 4D related applications, including motion retargeting, completion, and prediction can be achieved with high quality with our representation. Finally, we provide an ablation study to test the impact of each component in our framework on the reconstruction quality.

Dataset We use the CAPE dataset [38, 49] for training and evaluating, which is a dataset of 3D dynamic clothed humans containing 10 male and 5 female subjects wearing different types of outfits. More than 600 motion sequences of large pose variations are provided. In each sequence, the clothed body shapes are captured at 60 FPS along with corresponding meshes in the canonical pose and pose parameter of each frame. Overall, the dataset provides good di-

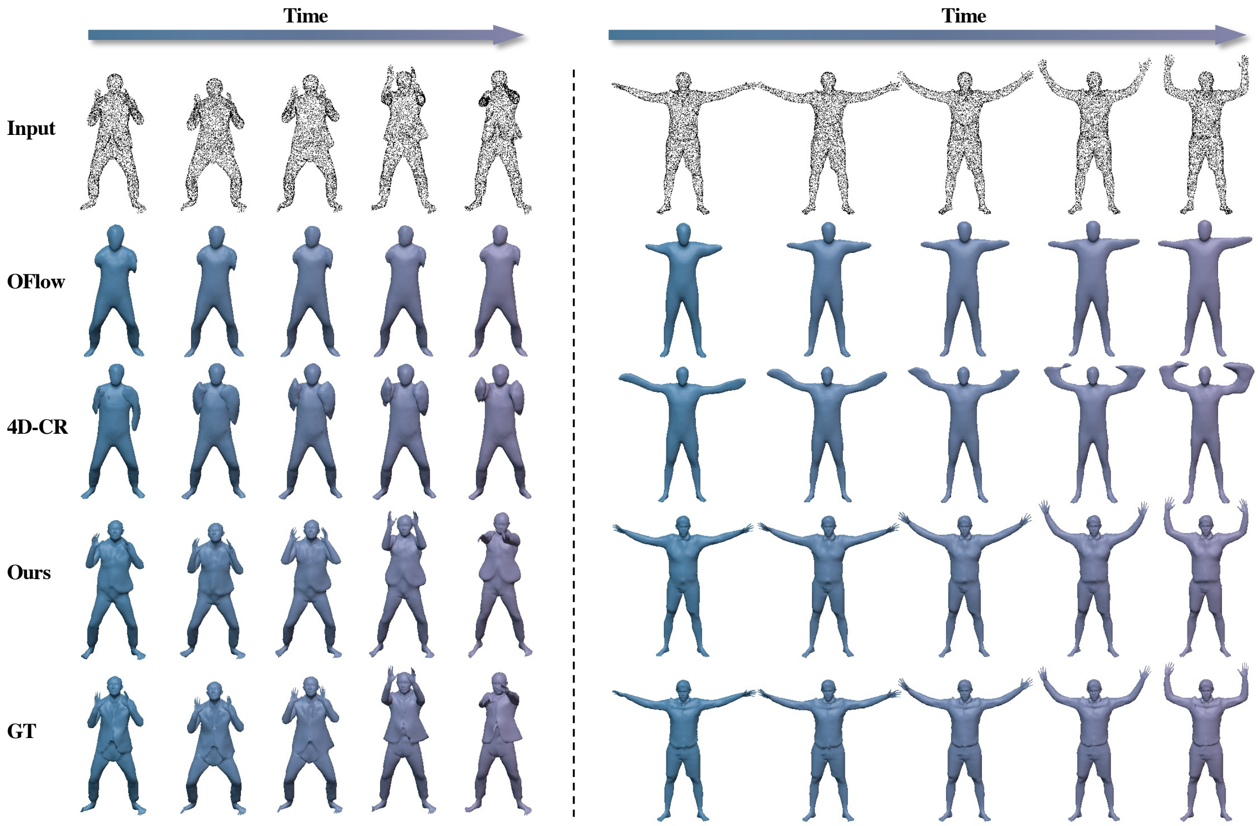


Figure 3. **4D Reconstruction.** Given the dense point cloud sequence (Row 1) uniformly sampled from the SMPL registered meshes, our method (Row 4) can reconstruct fine-grained meshes with accurate motion, while baseline methods (Row 2, 3) tend to be overly smooth and often have incomplete geometry, *e.g.* missing hands. We uniformly sample 5 frames (out of 30 frames) for visualization.

versity on both 3D geometry and motion. Following OFlow [44], we divide all sequences in CAPE into subsequences of 30 frames. We use subsequences from 488 motion sequences for training, and randomly sampled 2000 subsequences from the other 123 motion sequences for testing.

Implementation We use PyTorch to implement the model, and train with the Adam optimizer [31]. In the first stage, the learning rate is 10^{-4} with batch size 16. In the second stage, the initial learning rate is set to 10^{-4} and dropped to 10^{-5} after 200K iterations with batch size 4 due to the limitation of GPU memory. We use 4 NVIDIA GeForce RTX 2080Ti GPU cards.

Evaluation To measure the difference between the prediction and ground truth 3D shape, we use Chamfer Distance (CD) and Volumetric IoU (IoU) [43] for free-form geometry and Per Vertex Error (PVE) for SMPL registered shape. To measure the accuracy of motion, we use Procrustes-aligned mean per joint position error (PA-MPJPE), mean per joint position error (MPJPE), and acceleration error (mm/s^2) computed on 45 keypoints which include 24 joints and 21 keypoints on the face, feet and hands. Please refer to [30, 32, 43] for more details about these metrics. For temporal sequences, we take the mean score of all frames.

4.1. Representation Capability

We first show that our representation is capable of encoding and reconstructing human sequences with correct motion and geometry.

4D Reconstruction We compare to state-of-the-art 4D representation, Occupancy Flow (OFlow) [44] and 4D-CR [26], on mesh reconstruction from sampled point cloud inputs. As shown in Tab. 1 (I), our method significantly outperforms other methods on the 4D reconstruction accuracy. Qualitative results are shown in Fig. 3 for two temporal sequences. OFlow tends to produce incomplete geometry with missing hands, and the results from 4D-CR are overly smoothed, *e.g.* around face and hands. On the contrary, thanks to the human prior, our results are significantly better than others with complete geometry, correct motion, and rich details such as fingers, clothes, and hair. It is also worth noting that both OFlow and 4D-CR require point clouds with temporal correspondence as input, whereas our method can take unregistered point cloud sequences which is more convenient for many applications.

Shape and Motion Recovery We then study the performance of our motion model, which consists of a linear motion model and per-frame compensations recovered

I. Comparison with Previous 4D Representation Methods								
Methods	4D Reconstruction		Motion Retargeting		Motion Completion		Future Prediction	
	IoU ↑	CD ↓	IoU ↑	CD ↓	IoU ↑	CD ↓	IoU ↑	CD ↓
OFlow [44]	61.5%	0.199	30.7%	0.470	65.8%	0.181	58.8%	0.218
4D-CR [26]	62.9%	0.165	47.3%	0.296	76.6%	0.128	64.0%	0.200
Ours	**73.3%**	**0.093**	**70.7%**	**0.100**	**90.3%**	**0.031**	**71.7%**	**0.121**

II. Comparison with Human Body Estimation Methods (Forward)								
Methods	Shape and Motion Recovery				Motion Retargeting			
	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓
HMMR [30]	87.8	102.1	89.2	20.9	85.7	98.0	86.9	19.4
VIBE [32]	45.3	54.3	47.6	13.4	46.3	54.1	47.0	12.8
4D-CR-SMPL [26]	59.2	68.5	59.5	9.9	62.4	73.2	63.7	10.1
4D-CR-SMPL* [26]	49.8	57.7	49.8	8.9	52.2	59.6	51.6	8.7
Ours	**38.4**	**44.9**	**39.2**	**8.8**	**39.5**	**45.2**	**39.0**	**8.6**

III. Comparison with Human Body Estimation Methods (Backward)								
Methods	Motion Completion				Future Prediction			
	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓
HMMR [30]	146.5	141.6	148.3	48.7	148.4	142.9	146.9	48.3
Zhang *et al*. [66]	–	–	–	–	134.7	146.5	143.4	23.0
4D-CR-SMPL [26]	87.3	67.3	66.9	14.1	91.9	77.9	77.1	11.3
Ours	**53.8**	**42.7**	**41.7**	**9.4**	**73.1**	**62.8**	**59.7**	**11.2**

Table 1. **Comparison to SoTA methods on various tasks.** For evaluation, we adopt Volumetric IoU (IoU) and Chamfer Distance (CD) [43] for comparisons with free-form methods (block I), and several standard metrics following [30, 32] for SMPL-based methods (block II&III, the numbers are measured in *mm*). * denotes the input point cloud sequence has temporal correspondence.

from the auxiliary code. As baselines, we compare to SoTA video-based human shape and pose estimation methods HMMR [30] and VIBE [32]. Originally designed for color images input, we replace their image encoder with the point cloud encoder in our setup. Moreover, as an additional baseline, we extend 4D-CR [26] to an SMPL-based version by replacing their implicit occupancy decoder with the SMPL decoder so that it also benefits from the human prior. Since all of these methods produce only unclothed SMPL defined shapes, we disable our Shape-Comp network and use the output from the SMPL decoder for fair comparisons. All the baseline methods are retrained on our dataset. For extending 4D-CR, we train two models with registered point clouds as in their work (4D-CR-SMPL*) and unregistered point clouds like us (4D-CR-SMPL) respectively. The quantitative comparisons are shown in Tab. 1 (II). Our method achieves more accurate motion estimation (as measured at body keypoints by PA-MPJPE, MPJPE, acceleration error) and SMPL shape (as measured by PVE) than HMMR and VIBE. 4D-CR-SMPL performs relatively poorly when the input point cloud is unordered and gets much better once given tracked point clouds (4D-CR-SMPL*) but still performs worse than our method. We provide additional comparisons to the recent motion-based human body estimation method HuMoR [52] with the task of shape and motion recovery from point clouds via auto-decoding in Supp. Material.

4.2. Applications

Our representation can support various applications. Note that for all applications, the encoder and decoder are both fixed after training.

Motion Retargeting The goal of motion retargeting is to transfer the motion sequence from one subject to another. Traditional methods typically require manual works, *e.g.* provide correspondences between source and target identities [59], to fulfill such a task.

We achieve motion retargeting without any human intervention. Taking two point cloud sequences, one as the identity (I) and the other as the motion (M), we feed both into our compositional encoder to get SMPL parameters and latent codes for each $(c_s^I, c_p^I, c_m^I, c_a^I)$ and $(c_s^M, c_p^M, c_m^M, c_a^M)$. We then conduct the motion retargeting by using (c_s^I, c_p^M, c_m^M) for linear motion model, c_a^M for Motion-Comp network, and c_a^I for Shape-Comp network. Note that two c_a are used for Motion-Comp and Shape-Comp networks separately as they encode motion and shape information respectively.

For evaluation purpose, we randomly sampled 100 pairs of identity and motion sequences with $L = 30$ frames. We use the provided ground truth shape in the canonical pose and pose parameters provided by CAPE [38] dataset to generate motion retargeted ground truth sequences. We compare with free-form geometry based approach (OFlow and

4D-CR) on the full geometry (in Tab. 1 (I)) and SMPL based approach (HMMR, VIBE, 4D-CR-SMPL) on the unclothed body mesh from SMPL (in Tab. 1 (II)). Our method significantly outperforms OFlow and 4D-CR. As shown in a qualitative example in Fig. 4, our method produces much more complete motion retargeting results. Note how clothing details are successfully transferred, *e.g*. long trousers in the identity sequence compared to shorts in the motion sequence. Our method also outperforms all the human body estimation methods, showing that the compositional encoder is more effective in extracting correct information from inputs and facilitating motion recovering.

Motion Completion Our representation can also fulfill fitting tasks in the auto-decoding fashion, in which the SMPL parameters and latent codes are optimized to produce output similar to the observation. With this, our representation can perform motion completion, where the goal is to predict the missing data in a dynamic human sequence. For evaluation, we randomly choose 100 sequences with 30 frames from our test set. For each sequence, we randomly pick 15 frames as the observation, optimize the SMPL parameters and latent codes, reconstruct the full sequence, and then measure the geometry accuracy on the other 15 frames. Note that we use Chamfer loss with the additional prior terms borrowed from IPNet [8] on uniformly sampled points instead of PVE to simulate the case in real applications, where the observed meshes may not be registered.

Comparisons to free-form based methods and SMPL based methods are shown in Tab. 1 (I) and (III) respectively, and the qualitative results are in Supp. Material. Zhang *et al*. [66] uses a similar motion model to HMMR [30], so we only evaluate one of them. Overall, our method consistently outperforms all the other methods.

Moreover, we compare the robustness of auto-decoding based fitting using our Shape-Comp network against the naive CAPE decoder, and show the error of completion w.r.t the amount of random noise added to the observed frames in Fig. 6 (b). The error of our model is consistently lower than the naive CAPE decoder and deteriorates less with increasing noise. This is presumably because CAPE performs per-frame optimization, which may confuse the latent space if the gradients are not consistent from temporal frames, whereas we use GRU to model the temporal sequence for more robustness.

Last but not least, our model can also complete the temporal sequences from partial spatial observation. To show this, we generate one depth image per-frame from a camera rotating concurrently with the motion of the 3D shape, and run auto-decoding based fitting to complete the sequence. This can be also considered as a typical non-rigid fusion with known camera poses. We show the qualitative and quantitative comparisons to NPMs [45] in Supp. Material.

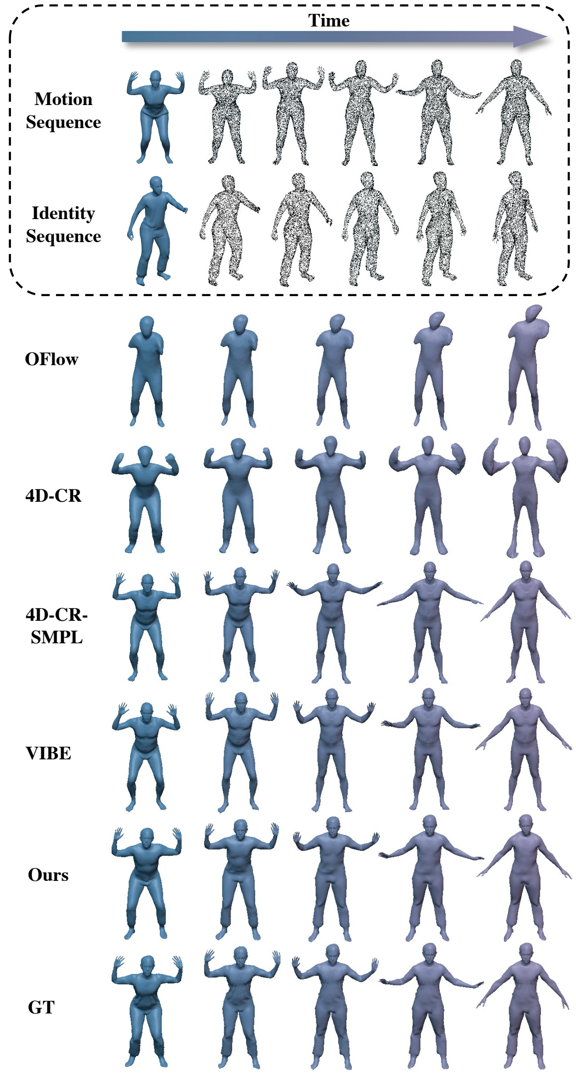


Figure 4. **Motion Retargeting.** Our goal is to transfer the human movements of the motion sequence (Row 1) to the people in the identity sequence (Row 2). We can accurately transfer the motion to the new identity and keep the original geometry details, *e.g*. garments and hairstyle, at the same time (Row 7). The free-form baselines (Row 3, 4) either fail due to shape and motion entanglement or produce more artifacts over time due to error accumulation. The SMPL-based baselines (Row 5, 6) also fulfill retargeting but are inaccurate, and they can only represent unclothed body.

Future Prediction Our representation also supports future prediction. Specifically, we run a fitting algorithm on the first 20 frames, generate the SMPL parameters and latent codes, and then reconstruct the full sequence to predict the 10 frames in the future. Tab. 1 (I, III) and Fig. 5 show the comparison to the previous methods. Again, we obtain significantly better performance than other 4D representation methods (OFlow and 4D-CR). When comparing only the motion accuracy using SMPL mesh with previous work on motion prediction (HMMR [30], Zhang *et al.* [66] and 4D-CR-SMPL [26]), our method still achieves better performance. Moreover, we empirically found that, though given pose prior terms during backward optimization, these baseline methods are more easily to produce unnatural poses than us, and predict unreasonable motions as shown in Fig. 5, possibly because our PCA-based motion model provides regularization and global context for output motions.

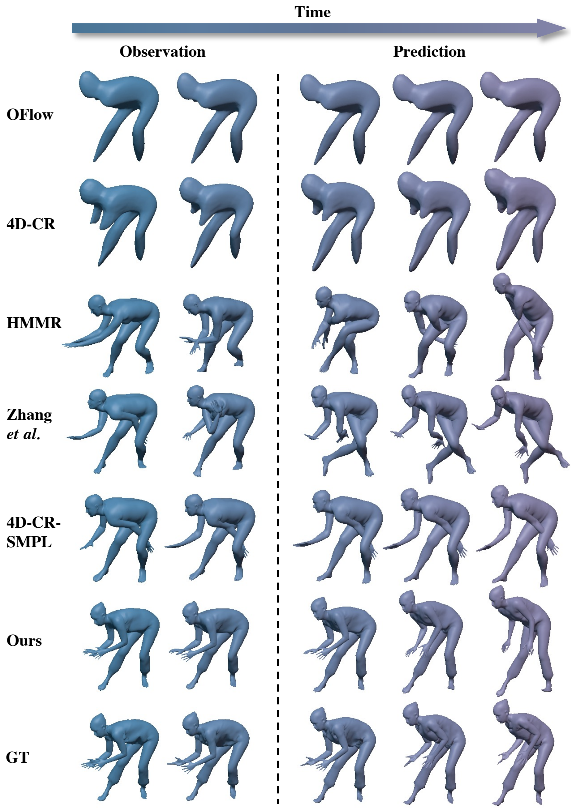


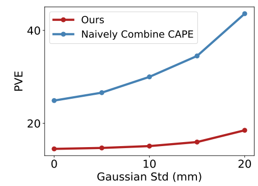
Figure 5. **Future Prediction.** We extrapolate 10 future temporal frames based on 20 past observed frames. The baseline methods (Row 1-5) either produce unsatisfactory geometry or stuck into the unnatural pose, while our method (Row 6) successfully keeps the movement trend and produces reasonable prediction of future motion. The meshes on the left and right are reconstructions of the observations and predictions for the future time steps, respectively.

4.3. Ablation Study

In this section, we perform an ablation study and show the quantitative results to demonstrate the effect of the major designs in our method. The visualization example can be found in Supp. Material.

Motion Model We first study the effect of the linear motion model and auxiliary code for motion recovery. We compare the ablation cases on the output of the SMPL decoder with the registered SMPL model on the ground truth mesh, which removes the free-form deformation and allows us to focus on the motion quality. In Fig. 6 (a), we show the performance of our model removing the linear motion model ("-LMM"), which directly updates the initial pose code, or removing Motion-Comp network ("-Motion c_a"), which only relies on the linear motion model (LMM) for motion

	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓
- GRU Enc.	49.6	57.0	49.6	10.6
- LMM	40.2	46.8	41.2	**8.8**
- Motion c_a	43.7	50.4	43.4	9.0
Full Model	**38.4**	**44.9**	**39.2**	**8.8**
- Shape c_a †	–	–	43.8	–
Full Model†	–	–	**42.0**	–



(a) Ablation Study (b) Noise Tolerance

Figure 6. (a) **Ablation study.** We verify the effectiveness of our temporal encoder by replacing the GRU with the modified PointNet used in [26, 44] (Row 1). Additionally, we remove the major modules in our framework in turn to demonstrate the effect of different components (Row 2, 3, 5). † denotes we compute the metrics with the ground truth clothed mesh. (b) **Robustness against noise.** The x-axis is the standard deviation of added Gaussian noise, and the y-axis is Per Vetex Error (PVE, lower is better).

recovery. In either case, the motion accuracy drops consistently as measured by all metrics, indicating the necessity of combining the prior model with learned compensations.

Shape Model We then verify if the auxiliary code helps to recover the detailed geometry. In Fig. 6 (a), we show the performance of the final mesh without and with auxiliary code-driven shape compensation (the last two rows), which is measured by PVE between the output mesh with the ground truth clothed mesh. The advantage of the shape compensation can also be found in Fig. 4 and 5, which show that the auxiliary code helps to improve the geometry details when comparing our results with the SMPL outputs from VIBE, HMMR or 4D-CR-SMPL.

Encoder Last but not least, we verify the effectiveness of our GRU-based temporal encoder. We replace our temporal encoder with the PointNet adopted in OFlow [44] and 4D-CR [26], and see a significant performance drop ("-GRU Enc."), which shows our GRU-based encoder helps to extract temporal information from the point cloud sequences without temporal correspondence.

5. Conclusion

This paper introduces H4D, a compact and compositional neural representation for 4D human captures, which combines the merits of both the prior-based and free-form solutions. A novel framework is designed for learning our representation, which encodes the input point cloud sequences into the SMPL parameters of shape and initial pose, and latent codes of motion and auxiliary information. Extensive experiments on 4D reconstruction, shape and motion recovery, motion retargeting, completion and prediction validate the efficacy of the proposed approach.

Acknowledgements

This work was supported in part by NSFC Project (62176061), Shanghai Municipal Science and Technology Major Project (2018SHZDZX01). The corresponding authors are Xiangyang Xue, and Yanwei Fu.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Representation learning and adversarial generation of 3d point clouds. *arXiv preprint arXiv:1707.02392*, 2(3):4, 2017. 2

[2] Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. *arXiv e-prints*, pages arXiv–2004, 2020. 2

[3] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7144–7153, 2019. 2

[4] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 2

[5] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 2

[6] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2293–2303, 2019. 2

[7] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018. 2

[8] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, aug 2020. 2, 7

[9] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 1, 2

[10] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *European Conference on Computer Vision*, pages 226–242. Springer, 2020. 2

[11] Rohan Chabra, Jan Eric Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[12] ChaoWen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *ICCV*, 2019. 2

[13] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018. 1, 2

[14] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6970–6981, 2020. 2

[15] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 2, 4

[16] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[17] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016. 2

[18] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 1, 2

[19] Philipp Erler, Paul Guerrero, Stefan Ohrhallinger, Niloy J. Mitra, and Michael Wimmer. Points2surf: Learning implicit surfaces from point clouds. In *ECCV*, 2020. 2

[20] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 2

[21] P Thomas Fletcher, Conglin Lu, Stephen M Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005, 2004. 2

[22] Rohit Girdhar, David F. Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 2

[23] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation. *arXiv preprint arXiv:1802.05384*, 2018. 2

[24] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 1, 2

[25] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*, pages 18–35. Springer, 2020. 4

[26] Boyan Jiang, Yinda Zhang, Xingkui Wei, Xiangyang Xue, and Yanwei Fu. Learning compositional representation for 4d captures with neural ode. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5340–5350, 2021. 1, 2, 3, 5, 6, 7, 8

[27] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas Funkhouser. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020. 2

[28] Ian T Jolliffe. Principal component analysis. *Technometrics*, 45(3):276, 2003. 2, 3

[29] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 2

[30] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 1, 2, 5, 6, 7

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[32] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 1, 2, 5, 6

[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/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 1, 2

[34] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017. 1, 2

[35] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *2019 International Conference on 3D Vision (3DV)*, pages 643–653. IEEE, 2019. 2

[36] Yiyi Liao, Simon Donne, and Andreas Geiger. Deep marching cubes: Learning explicit surface representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2916–2925, 2018. 2

[37] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1, 2, 3

[38] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4, 6

[39] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019. 2

[40] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017. 2

[41] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera. *arXiv preprint arXiv:1907.00837*, 2019. 4

[42] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 1, 2

[43] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 5, 6

[44] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5379–5389, 2019. 1, 2, 3, 5, 6, 8

[45] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12695–12705, 2021. 7

[46] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[47] 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. 2

[48] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018. 1, 2

[49] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):1–15, 2017. 4

[50] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum pointnets for 3d object detection from RGB-D data. In *CVPR*, 2018. 2

[51] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 2, 3

[52] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11488–11499, 2021. 6

[53] Davis Rempe, Tolga Birdal, Yongheng Zhao, Zan Gojcic, Srinath Sridhar, and Leonidas J Guibas. Caspr: Learning canonical spatiotemporal point cloud representations. *Ad-*

vances in Neural Information Processing Systems, 33, 2020. 2

[54] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017. 2

[55] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2

[56] 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. 2

[57] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2886–2897, 2021. 2

[58] 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. 1

[59] Robert W Sumner and Jovan Popović. Deformation transfer for triangle meshes. *ACM Transactions on graphics (TOG)*, 23(3):399–405, 2004. 6

[60] Pavel Tokmakov, Yu-Xiong Wang, and Martial Hebert. Learning compositional representations for few-shot recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6372–6381, 2019. 2

[61] 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(2-3):157–177, 2006. 2, 3

[62] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018. 2

[63] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions on Graphics (TOG)*, 36(4):72, 2017. 2

[64] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3d human mesh registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7639–7648, 2021. 2

[65] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5908–5917, 2019. 2

[66] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7114–7123, 2019. 1, 2, 4, 6, 7

[67] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2