SynBody: Synthetic Dataset with Layered Human Models for 3D Human Perception and Modeling

Zhitao Yang\textsuperscript{1,*} Zhongang Cai\textsuperscript{1,2,3,*} Haiyi Mei\textsuperscript{1,*} Shuai Liu\textsuperscript{2,*} Zhaoxi Chen\textsuperscript{3,*} Weiye Xiao\textsuperscript{1} Yukun Wei\textsuperscript{1} Zhongfei Qing\textsuperscript{1} Chen Wei\textsuperscript{1} Bo Dai\textsuperscript{2} Wayne Wu\textsuperscript{2} Chen Qian\textsuperscript{1} Dahua Lin\textsuperscript{4} Ziwei Liu\textsuperscript{3,†} Lei Yang\textsuperscript{1,2,†}

\textsuperscript{1}SenseTime Research \textsuperscript{2}Shanghai AI Laboratory \textsuperscript{3}S-Lab, Nanyang Technological University \textsuperscript{4}The Chinese University of Hong Kong

\textsuperscript{*}Equal Contribution \textsuperscript{†}Corresponding Author

https://synbody.github.io/

Figure 1: SynBody is a large-scale synthetic dataset with a massive number of subjects and high-quality annotations. It supports various research topics, including human pose and shape estimation (HPS) and novel view synthesis for human (Human NeRF).

Abstract

Synthetic data has emerged as a promising source for 3D human research as it offers low-cost access to large-scale human datasets. To advance the diversity and annotation quality of human models, we introduce a new synthetic dataset, SynBody, with three appealing features: 1) a clothed parametric human model that can generate a diverse range of subjects; 2) the layered human representation that naturally offers high-quality 3D annotations to support multiple tasks; 3) a scalable system for producing realistic data to facilitate real-world tasks. The dataset comprises 1.2M images with corresponding accurate 3D annotations, covering 10,000 human body models, 1,187 actions, and various viewpoints. The dataset includes two subsets for human pose and shape estimation as well as human neural rendering. Extensive experiments on SynBody indicate that it substantially enhances both SMPL and SMPL-X estimation. Furthermore, the incorporation of layered annotations offers a valuable training resource for investigating the Human Neural Radiance Fields (NeRF).
1. Introduction

The fields of 3D human perception [21, 23–25, 37, 53] and human reconstruction [13, 14, 26, 40, 41] have become increasingly important, but the lack of available data has limited their development. Collecting real human data on a large scale is challenging due to privacy concerns and time constraints. Therefore, exploring the use of synthetic human datasets has become a critical avenue of research.

Despite the great potential, existing synthetic datasets [5, 8, 38, 47] suffer from limitations such as the number of available human models and the quality of annotations. The main reason lies in that synthetic human datasets rely on real scans for rendering, which poses three key obstacles. Firstly, it is challenging to expand the types of body shapes, poses, and clothing available in the dataset. Secondly, as the human models are scanned with clothing, the 3D annotations obtained through fitting are prone to errors. Thirdly, it is difficult to obtain annotations of body and clothing separately. To address these issues, we develop a new synthetic dataset termed SynBody. The dataset includes 1.2 million frames with corresponding ground-truth 3D human body annotations. It covers 10,000 human body models, 1,187 motions, and 26,960 video clips with 2.7M SMPL/SMPL-X annotations.

At the heart of SynBody is the layered parametric human model, which constructs the clothed human model in a bottom-up manner. SMPL-X [39] is a widely used parametric human model, capable of sampling human models with various body shapes. However, it lacks the ability to model clothing, limiting its applicability when synthesizing realistic human models. To overcome this limitation, we introduce SMPL-XL, a parametric human model based on SMPL-X in a layered representation. SMPL-XL enriches the SMPL-X model in three aspects: (1) Hair system: adding hair and beards to the FLAME [27] model, with 32 types of hair and 13 types of beards; (2) Garment and accessories: adding procedural clothes to the SMPL-X body, including coats, shirts, pants, skirts, shoes, and glasses; (3) Texture: in addition to adding rich geometry, SMPL-XL also adds rich textures for sampling various skin colors and clothing textures.

The designed SMPL-XL is capable of automatically generating a large number of human models with high-quality annotations. We therefore generate 10,000 clothed human models by sampling various body shapes, clothing styles, hairstyles, accessories, and textures. Notably, the use of the SMPL-X model as the base body model guarantees that the parametric human annotations are always accurate, obviating the need for the necessity for annotations through fitting. Furthermore, as the clothes are explicitly attached to the surface of the human body, layered annotations for body and clothes are available.

To generate a large-scale dataset with high diversity and high-quality annotations, we design a scalable and automatic system to render images and annotations. We first animate the 10,000 dressed human models by retargeting motions from a large motion library [31]. Subsequently, we design an algorithm to place human models in the scene without piercing. Multiple cameras are then placed by evaluating self-occlusion, inter-occlusion, and view diversity, and the rendering module renders the assets into images with corresponding annotations.

With SynBody, we launch two tracks that support human pose and shape estimation and human neural rendering, respectively. Experiments show that SynBody is more effective than AGORA under the same amount of training data for human pose and shape estimation. With diverse and large-scale training data, SynBody achieves significant performance gains on both SMPL and SMPL-X estimation. In terms of human neural rendering using neural radiance fields (i.e., NeRF [34]), benchmarking existing approaches on SynBody shows that it has comparable performance as real human data. Furthermore, with the layered annotations which offer accurate SMPL parameters, we observe that current human NeRF approaches are sensitive to the accuracy of estimated SMPL.

In summary, SynBody is a large-scale synthetic dataset for human perception and modeling, with three main contributions: (1) It constructs clothed subjects and samples 10,000 animatable subjects, which is an order of magnitude higher than existing datasets. (2) The clothed subjects are constructed with an explicit cloth model, thus it provides layered 3D annotations of the human body and clothing, which are not available in previous datasets. (3) Experiments on SynBody achieve promising results on both human perception and modeling, emphasizing the importance of diversity and annotation quality for downstream tasks.

2. Related Works

Human Parametric Models. Several 3D human parametric models, such as SMPL [30], SMPL-X [39], and GHUM [51], have been developed to generate 3D human meshes from parameters that represent the human pose and shape using linear blend skinning. SMPL-X [39] extends SMPL [30] by combining FLAME [27] and MANO [43] for the head and hands, respectively, and is trained on a large number of real scans to provide a strong basis for shape variations. However, SMPL-X only produces naked body meshes, and we aim to enhance its realism by building a layered parametric model that includes hair, clothes, and accessories. The proposed model leverages the shape basis of SMPL-X while providing realistic dressed human meshes.

Human Pose and Shape Estimation. Several methods [21, 23–25, 37, 49, 52–54] have been proposed to estimate 3D human pose and shape parameters. HMR [21] directly regresses these parameters in an end-to-end manner, while
Table 1: Comparisons of 3d human dataset. We compare SynBody with existing datasets. We divide datasets into three types: real (R), synthetic (S), and mixed (M). SynBody constructs 10,000 animatable subjects, which is an order of magnitude higher than any existing datasets and brings competitive scale and diversity. “ITW” stands for “In-the-Wild” in the table.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>ITW</th>
<th>Video</th>
<th>#Views</th>
<th>#SMPL</th>
<th>#Seq</th>
<th>#Subj.</th>
<th>#Motions</th>
<th>GT format</th>
</tr>
</thead>
<tbody>
<tr>
<td>HumanEva [45]</td>
<td>R</td>
<td>-</td>
<td>✓</td>
<td>4/7</td>
<td>NA</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>3DJ</td>
</tr>
<tr>
<td>Human3.6M [15]</td>
<td>R</td>
<td>-</td>
<td>✓</td>
<td>4</td>
<td>312K</td>
<td>839</td>
<td>11</td>
<td>15</td>
<td>3DJ, SMPL</td>
</tr>
<tr>
<td>MPI-INF-3DHP [32]</td>
<td>M</td>
<td>✓</td>
<td>✓</td>
<td>14</td>
<td>96K</td>
<td>16</td>
<td>8</td>
<td>8</td>
<td>3DJ</td>
</tr>
<tr>
<td>3DPW [48]</td>
<td>R</td>
<td>✓</td>
<td>✓</td>
<td>1</td>
<td>32K</td>
<td>60</td>
<td>18</td>
<td>*</td>
<td>SMPL</td>
</tr>
<tr>
<td>Panoptic Studio [20]</td>
<td>R</td>
<td>✓</td>
<td>✓</td>
<td>480</td>
<td>736K</td>
<td>480</td>
<td>~100</td>
<td>*</td>
<td>3DJ</td>
</tr>
<tr>
<td>EFT [19]</td>
<td>R</td>
<td>✓</td>
<td>-</td>
<td>1</td>
<td>129K</td>
<td>NA</td>
<td>Many</td>
<td>NA</td>
<td>SMPL</td>
</tr>
<tr>
<td>ZJU-MoCap [41]</td>
<td>R</td>
<td>✓</td>
<td>✓</td>
<td>21</td>
<td>180K</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>SMPL, mask</td>
</tr>
<tr>
<td>SURREAL [47]</td>
<td>S</td>
<td>✓</td>
<td>✓</td>
<td>1</td>
<td>6.5M</td>
<td>NA</td>
<td>145</td>
<td>2K</td>
<td>SMPL</td>
</tr>
<tr>
<td>AGORA [38]</td>
<td>S</td>
<td>✓</td>
<td>✓</td>
<td>1</td>
<td>173K</td>
<td>NA</td>
<td>&gt;350</td>
<td>NA</td>
<td>SMPL, SMPL-X, mask</td>
</tr>
<tr>
<td>HSPACE [5]</td>
<td>S</td>
<td>✓</td>
<td>✓</td>
<td>5</td>
<td>-</td>
<td>NA</td>
<td>100×16</td>
<td>100</td>
<td>GHUM/L, mask</td>
</tr>
<tr>
<td>GTA-Human [8]</td>
<td>S</td>
<td>✓</td>
<td>✓</td>
<td>1</td>
<td>1.4M</td>
<td>20K</td>
<td>&gt;600</td>
<td>20K</td>
<td>SMPL</td>
</tr>
<tr>
<td>SynBody</td>
<td>S</td>
<td>✓</td>
<td>✓</td>
<td>4</td>
<td>2.7M</td>
<td>27K</td>
<td>10,000</td>
<td>1,187</td>
<td>SMPL, SMPL-X, mask</td>
</tr>
</tbody>
</table>

SPIN uses an optimization step [25] to guide the learning process towards pseudo-3D labels. PARE [24] employs part attention to tackle occlusion, and VIBE [23] leverages temporal information in videos for SMPL estimation. Apart from building low-cost data collection solutions [7], synthetic datasets have become a promising alternative to efficiently scale up data. SURREAL [47] renders textured SMPL body models in real-image backgrounds but does not account for cloth geometry, resulting in unrealistic subjects. AGORA [38] renders real human scans in a virtual world and provides high-quality synthetic data for image-based approaches. HSPACE [5] places animated human models in various scenes to provide training data for video-based methods, and increases the variation of human shape via refitting. GTA-Human [8] captures videos and optimizes corresponding SMPL annotations from video games. However, the diversity of subjects in current datasets is limited by either body shapes or cloth types.

Expressive Human Pose and Shape Estimation. As face and hand are also crucial for human perception, some efforts [35,44,55] have been made to whole-body human pose and shape estimation. ExPose [9] introduces three experts to predict parameters for body, hand, and face, and merge them in a copy-paste strategy. Hand4Whole [35] further improves the prediction of wrist and finger poses by leveraging selected hand joint features. Predicting both hands and faces makes the dataset much more difficult to obtain than just predicting the body. In order to increase the diversity of real data, AGORA [38] provides image-based SMPL-X annotations by fitting SMPL-X on the scanned human models. While a very recent work, BEDLAM [6], also enhances the SMPL-X model with clothing, hair, and changes in skin tones. However, it employs a manual approach for creating garments and utilizes pre-made skin textures, making it challenging to generate a large number of human body models. In contrast, our body model construction process is fully automated, encompassing clothing, hair, textures, and more. This enables easier scalability to generate human body models on a large scale.

Human NeRF. NeRF [34] has demonstrated impressive photo-realistic view synthesis by learning implicit fields of density and color. Yet, human motions are more challenging to learn due to dynamic deformation fields. NeuralBody [41] incorporates prior from a statistical body template to learn dynamic sequence, while AnimatableNeRF [40] proposes to reconstruct an animatable human model that generalizes to new poses. Furthermore, NHP [26] and KeypointNeRF [33] achieve generalizability to unseen identities and poses. Several datasets have been adapted to study human NeRF. ZJU-MoCap [41] captures 9 human subjects with 21 synchronized cameras, providing fitted human body model parameters as well as the foreground mask. Human3.6M [16] collects 11 human subjects with 4 cameras, using a marker-based motion capture system. A-NeRF [46] generates a synthetic dataset using SURREAL [47] to study factors that affect visual quality.

While the aforementioned tasks are interrelated and often draw upon similar datasets, as shown in Table 1, datasets are limited in two key aspects. Firstly, obtaining real human models is challenging, which restricts the scale of these datasets. Secondly, 3D annotations are typically acquired through optimization, which introduces errors and cannot provide layered annotations. In contrast, with the designed SMPL-XL model, SynBody provides 10,000 subjects with diverse body shapes and clothing, along with layered annotations that include accurate SMPL and SMPL-X. Built on top of this layered human model, SynBody comprises two subsets for human mesh recovery and human NeRF.
tively, and features the same 10,000 diverse subjects.

3. Synthetic Data Generation System

![Diagram of the synthetic data generation system](image)

Figure 2: Synthetic data generation system. It consists of 4 components: 1) human model creation to generate layered human models, 2) motion retargeting to drive human models, 3) scene composition to place actors and cameras, and 4) rendering and outputs to generate multi-modal dataset.

Similar to movie production [4], our system comprises 4 components (depicted in Figure 2): (1) a layered parametric human model creation service, a scalable process to generate layered human models, (2) a motion retargeting module to apply motions from various sources to layered human models, (3) a scene composition module to place 3D actors and objects into a 3D scene, setting up cameras, (4) a 3D rendering engine and a multi-modal data annotation generator. Our infrastructure enables the generation of high-quality synthetic data for various computer vision tasks.

3.1. Layered Parametric Human Model

Parametric human models like SMPL-X [39] provide the ability to create rigged body models with various body shapes. However, the lack of available textures limits its application in data generation. We designed a module to perform an automatic process that combines SMPL-X with procedural garments and accessories, hair system and textures, producing realistic and diverse body models.

**Body shape.** SMPL-X [39] body model has a 3D mesh whose vertex locations are controlled by parameters for pose $\theta$, shape $\beta$, and facial expression $\psi$. By modifying shape $\beta$, we obtain 3D meshes of various human heights and weights. And by alternating pose $\theta$, meshes can be driven to perform various poses.

**Garment Model.** Our garment is generated as a separate layer on top of the body. Following the industrial garment-making workflow, we designed garment patterns of various types. Notice that different parts of garment pieces are connected with sewing lines, e.g., the red line in Figure 4 (a). Then, we stitch patterns onto the body at T-Pose utilizing a physical simulator [1]. Specifically, we first manually move the garment pieces to roughly align them with the body. During simulation, vertices between two ends of the seam gradually shrink until they completely pooled together. Figure 4 (b) demonstrates the final draped garment. For garment animation, we bind every vertex in the garment to the closest point in the body. Then, the skinning weight and blend shape of the body mesh are assigned to garment vertices. This makes our garment model easily integrated with existing skeletal pipelines with little computational overhead.

**Particle hair system.** Our method uses a prefabricated particle system to generate realistic hair on a head-shaped mesh. To achieve this, a template of hairstyle or facial hair $T_{\text{hair}}$ would attach to the vertices on the mesh which are marked with different areas $V_{\text{hair}}$ (including fringe, top, temporal bone, occipital bone, and bottom area). Designers draw multiple sets of guidelines $L_{\text{guide}}$ with varying shapes on different areas. Each set of guidelines comprises a collection of Bezier curves that are utilized to accurately constrain the flow of hair strands. Furthermore, the shape of the hair strands can be adjusted by length $L_{\text{length}}$ and curliness $L_{\text{curliness}}$. The entire hair system is composed of multiple sub-particle systems $\{V_{\text{hair}}, L_{\text{guide}}, L_{\text{length}}, L_{\text{curliness}}\}$.

**Accessories.** We also add template accessories $T_{\text{accessories}}$ to our model, such as glasses, shoes, hats, and headphones. All accessories are pre-assembled on an SMPL-X template with a uniform body shape. These accessories can be transferred to models with the same topology, ensuring consistent deformation across all models in accordance with changes in the shape $\beta$ of SMPL-X. The corresponding bone weights of the human body model’s vertices are transferred to the nearest accessories’ vertices, ensuring that these accessories are correctly driven by the armature.

**Procedural texture.** The procedural garment textures $T_{\text{procedural}} = \{T_{\text{pattern}}, T_{\text{decals}}, T_{\text{bump}}, P_{\text{mapping}}\}$ used in clothing are composed of multiple pre-set textures by alpha blending as demonstrated in Figure 4 (c). The pattern texture $T_{\text{pattern}}$ and decals texture $T_{\text{decals}}$ serve as layered masks to elevate the style and color, while the bump texture $T_{\text{bump}}$ functions as a height map, adding detailed normal information to the clothing’s surface. The mapping parameters $P_{\text{mapping}}$ control the coordination of all the textures in the UV space. Besides, to procedurally create body mesh textures, we build a similar texture template SMPL-X, in which we pre-draw layered masks to separate different body features, such as skin, lips, and freckles, allowing for color adjustment and blending in specific regions as in Figure 3.

Combining all the elements above, as shown in Figure 5, we obtain human body models with the same high quality as RenderPeople [2]. Besides, it is challenging to expand the types of body shape and clothing available for real scans, but our model can be easily scaled by randomly sampling each component.

3.2. Motion Retargeting

**Retargeting of skeletal animations.** Our motion retargeting module allows for the transfer of motion data from vari-
Figure 3: Layered parametric human model creation: (a) Generation of a naked human model using parametric body model [39], procedural body texture colors, and particle hair System; (b) Integration of garments and accessories onto the naked body model.

Figure 4: Demonstrations of Garment and Procedural Texture. (a) garment pattern with sewing lines, (b) simulated garment under the body in canonical space, (c) procedural textured garment whose style and color are elevated by layered masks provided by patterns and decals.

Figure 5: A Demonstration of comparison between commonly used human models in existing synthetic datasets [5, 8, 38, 47] and our SMPL-XL. We obtain high-quality models equal to RenderPeople [2], both are much better than SURREAL [47], and our model has the capability of scaling up easily by sampling various assets and body shapes.
The optimization process consists of two steps. First, we fit the shape $\beta$ of all human models under the T-pose to its corresponding SMPL-X. Secondly, for each sequence, it is initialized with fitted $\beta$ and its original pose. We fix $\beta$ while fitting the body pose $\theta$ for SMPL.

### 3.3. Scene Placement

To place $N_p$ subjects in a large scene with $N_o$ objects, we primarily follow three principles: standing on the ground, avoiding human-object and human-human penetration. To prevent subjects from floating in the air, the root position of a subject should align with the ground height. A sequential decision-making approach is used to find a suitable position for each subject, i.e., placing one subject at a time. An object is represented by $o_i = \{x_i, y_i, l_i, w_i\} \in \mathbb{R}^4$, where $\{x_i, y_i\}$ is the center of the object’s axis-aligned bounding box projected onto the ground with length $l_i$ and width $w_i$. Different from static objects, to avoid collisions between moving subjects at any one time, a subject with specified body shape and motion is simplified to $q_i = \{x_i, y_i, l_i, w_i\} \in \mathbb{R}^4$, which is the smallest axis-aligned box that envelops all bounding boxes across frames.

To avoid human-object penetration and potential human-human collision, the solution $p_j^* = \{x_i^*, y_i^*, l_i, w_i\}$ of a character with the shape of $\{l_i, w_i\}$ should satisfy:

$$I(\{x_i^*, y_i^*, l_i, w_i\}, o_j) = 0, \quad j = 1, \ldots, N_o \tag{3}$$

$$I(\{x_i^*, y_i^*, l_i, w_i\}, q_k) = 0, \quad k = 1, \ldots, N_p \tag{4}$$

where $I(box_1, box_2)$ denotes the overlapping area between two boxes and $N_p$ is the number of subjects already placed in the scene. Besides, distance constraint is used to prevent subjects from excessive dispersal. The problem is solved by grid search. Typically multiple solutions are available and we randomly sample one each time.

### 3.4. Camera Placement

Once the positions of all subjects have been organized, the next task involves placing $N_c$ cameras in suitable locations. To ensure that cameras would not be placed inside any objects or subjects, the 3D version of Eq. (3) and Eq. (4) are applied. In addition, we evaluate the suitability of a candidate camera by the following metrics: the distance from the camera to the subjects, the camera’s pitch angle, and the degree of occlusion of each subject in the camera’s view.

The distance denoted as $L$ from the mean position of all subjects $\bar{p}$ to the camera is restricted. $L_{\text{max}}$ is set to prevent an unreasonably small proportion of subjects in the image. To control the visibility of subjects, $L_{\text{min}}$ is defined as $\frac{\lambda}{\sum_{i=1}^{N_c} \alpha} \max_i ||p_i - \bar{p}||_2$, where $i = 1, \ldots, N_c$, $\alpha$ is the field of view of the camera, and $\lambda$ is a hyperparameter that determines the probability of all subjects being within the camera’s view. To estimate the degree of occlusion for each subject, rays are randomly and uniformly cast from the camera to each subject’s body, and the percentage of blocked rays by other objects is determined. More details can be found in the Sup. Mat.

### 3.5. Rendering and Annotations

Using a high-quality rendering pipeline in Unreal Engine 5, SynBody is rendered in multi-view with large 3D environments from the Unreal Marketplace for rich background information and dynamic lighting. Leveraging the G-buffer [12], our system simultaneously generates photo-realistic RGB images and annotations, along with accurate ground truth for segmentation masks, optical flow, depth maps, normal maps, and other ground-truth labels.

### 4. SynBody Dataset

#### 4.1. Dataset Statistics

SynBody comprises 10,000 unique SMPL-XL models randomly created with different body shapes and genders. Each model is then combined with the following assets: (1) hairstyles $T_{\text{hair}}$ sampled from 45 particle hairs; (2) garments $G_{\text{tmp}}$ sampled from 68 clothing models containing multiple outfits; (3) procedural texture $T_{\text{procedural}}$ generated by sampling $T_{\text{pattern}}$, $T_{\text{decals}}$, and $T_{\text{bump}}$ from 1,038 template textures, and color values are randomly sampled; (4) accessories $T_{\text{accessories}}$ sampled from 46 template assets. To ensure the validity of the motions, we select a subset from AMASS [31]. Based on the BABEL annotations [42], we excluded interactive motions and non-ground motions (e.g., swim), as well as filtered out motions with a duration shorter than 2 seconds, leading to a subset with 1,187 motions. For each sequence, we randomly selected 4 human models, and each model is assigned a randomly chosen motion lasting between 2 and 10 seconds. For motions exceeding 2 seconds, we randomly extract a 2-second segment to ensure that each video clip has a length of 60 frames. To enhance the diversity of viewpoints, 4 view positions are generated for each sequence, where each one is randomly sampled from the surface of semi-spheres with varying radii. The system generates 26,960 sequences and 1.2M images with annotations, utilizing 6 vast and realistic scenes created by professional artists. 2.7M SMPL/SMPM-X annotations are provided, excluding highly occluded subjects. More details can be found in the Sup. Mat.

#### 4.2. Human Pose and Shape Estimation

With 2.7 million SMPLs, we leverage a pretrained regressor to extract 3D keypoints and then project them onto 2D space. Following standard practice in top-down human mesh recovery methods, we generate bounding boxes from the resulting 2D keypoints.
4.3. Human Neural Rendering

Given the flexibility of our pipeline, we render a total of 100 multi-view sequences with diverse motions and appearances for benchmarking human NeRFs. All sequences have a length of 300 frames with a resolution of $1024 \times 1024$, where the motion sequence is randomly sampled from AMASS [31]. Each sequence contains 8 views whose camera positions have uniformly distributed azimuth angles around the human body. We use RGB renderings, binary foreground masks, camera parameters, and SMPL parameters for the benchmark. Thanks to our layered design, we can offer accurate SMPL parameters for human NeRFs, which acts as an important prior for a majority of methods, while keeping the diversity in clothes and motions.

5. Experiment

In this section, we study the usefulness of SynBody for two popular research directions: human pose and shape estimation and human NeRF.

5.1. Human Pose and Shape Estimation

Estimating 3D humans represented by SMPL (body-only) and SMPL-X (body, hands, and face) parameters from monocular 2D input has gained substantial attention. SynBody emerges as a scalable and effective solution to tackle the scarcity of paired data in these fields.

Parametric Human Models. SMPL [30] represents high-dimension human mesh $M(\theta, \beta) \in \mathbb{R}^{6890 \times 3}$ as low-dimension pose parameters $\theta \in \mathbb{R}^{72}$ and shape parameters $\beta \in \mathbb{R}^{10}$. In addition to SMPL, SMPL-X [39] further consists of additional left and right hand pose parameters $(\theta_{lh}, \theta_{rh} \in \mathbb{R}^{15 \times 3})$, jaw joint rotation $(\theta_{j} \in \mathbb{R}^{3})$ and face expression parameters $(\phi_{f} \in \mathbb{R}^{10})$ for an expressive human representation.

Benchmarks. We conduct experiments on two mainstream benchmarks: 3DPW [48] and AGORA [38]. As a widely used benchmark for in-the-wild evaluation, 3DPW encompasses diverse data collected through a mobile phone cam-

 era, paired with SMPL annotations. As for AGORA, it is a recent synthetic dataset that features challenging scenes with person-person occlusion. AGORA provides both SMPL and SMPL-X annotations.

Evaluation Metrics. To evaluate the quality of predicted parametric human models, we employ the standard metrics: $\text{MPJPE}$ (Mean Per Joint Position Error), which is calculated as the average $L_2$ distance over 3D keypoints regressed from parametric models; $\text{PA-MPJPE}$ (Procrustes Aligned Mean Per Joint Position Error), which is MPJPE with applying Procrustes Alignment [11] on the predicted keypoints to match the ground truth before error computation; $\text{PVE}$ (Per Vertex Error), which is the average $L_2$ distance between predicted and ground truth mesh vertices.

5.1.1 SMPL Estimation

Methods. To gauge the usefulness of SynBody, we conduct experiments with four milestone works: HMR [21], SPIN [25], and PARE [24] are image-based methods and VIBE [23] is a video-based method. We follow the hyper-parameter configurations provided in MMHuman3D [10].

Training Data. In the following experiments, we refer to the standard baskets of real datasets for SMPL estimation as "R". For HMR and SPIN, R consists of H36M [16], MPI-INF-3DHP [32], LSP [17], LSPE [18], MPII [3] and MSCOCO [29], whereas in VIBE, R consists of MPI-INF-3DHP and InstaVariety [22]. "S" denotes SynBody data with 2.7M training instances with SMPL annotations.

Main Results. We collate experiment results on 3DPW in Table 2. Each baseline is fine-tuned with a mixture of real datasets (R) and SynBody (S), leading to significant performance gains across different baseline models, even on strong baseline PARE [24] (2.81 mm and 1.41 mm improvements on MPJPE and PVE). Notably, SynBody can be used to train video-based methods such as VIBE as it contains video sequences instead of static images. Table 3 illustrates that training with SynBody exhibits remarkable improvements across methods, when compared with...
Figure 6: Impact of the amount of SynBody data. The horizontal axis represents the amount of SynBody data that is varied by multiples of real data. The baseline model is HMR and tested on the 3DPW test set.

Table 4: Comparison between SynBody-100K and AGORA on 3DPW test set. “A” means AGORA, and “S100K” means SynBody-100K which is a subset of SynBody datasets with over 100K SMPL annotations. For a fair comparison, the total number of SMPL annotations is close to that of AGORA.

<table>
<thead>
<tr>
<th>Method</th>
<th>Datasets</th>
<th>MPJPE</th>
<th>PA-MPJPE</th>
<th>PVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMR</td>
<td>R + A</td>
<td>101.61</td>
<td>57.85</td>
<td>123.82</td>
</tr>
<tr>
<td>HMR</td>
<td>R + S100K</td>
<td>95.28</td>
<td>57.68</td>
<td>119.18</td>
</tr>
<tr>
<td>SPIN</td>
<td>R + A</td>
<td>88.44</td>
<td>54.97</td>
<td>110.35</td>
</tr>
<tr>
<td>SPIN</td>
<td>R + S100K</td>
<td>85.52</td>
<td>54.12</td>
<td>105.45</td>
</tr>
<tr>
<td>PARE</td>
<td>R + A</td>
<td>85.34</td>
<td>48.39</td>
<td>109.77</td>
</tr>
<tr>
<td>PARE</td>
<td>R + S100K</td>
<td>79.42</td>
<td>47.80</td>
<td>102.45</td>
</tr>
</tbody>
</table>

Table 5: Effectiveness of SMPL-XL. R: real datasets (MSCOCO, MPI-INF-3DHP, Human3.6M, LSP, LSPET, MPII). S100K: downsampled SynBody with 100K instances. *: SURREAL-style.

<table>
<thead>
<tr>
<th>Method</th>
<th>Datasets</th>
<th>MPJPE</th>
<th>PA-MPJPE</th>
<th>PVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARE</td>
<td>R</td>
<td>81.8</td>
<td>49.4</td>
<td>105.3</td>
</tr>
<tr>
<td>PARE</td>
<td>R+S100K*</td>
<td>82.2</td>
<td>49.7</td>
<td>105.6</td>
</tr>
<tr>
<td>PARE</td>
<td>R+S100K</td>
<td>79.4</td>
<td>47.8</td>
<td>102.5</td>
</tr>
</tbody>
</table>

Table 6: SMPL-X estimation with OSX [28] as the baseline on AGORA validation set (AGORA-val) [38] and 3DPW test set [28]. R: real datasets (MSCOCO, MPII, and Human3.6M). S: SynBody. AGORA-val uses PVE (mm) whereas 3DPW uses MPJPE (mm) and PA-MPJPE (mm). Note the original OSX uses the SMPL head for 3DPW, which we modify to the SMPL-X head.

<table>
<thead>
<tr>
<th>Method</th>
<th>Datasets</th>
<th>AGORA</th>
<th>3DPW (Body)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>Hands</td>
</tr>
<tr>
<td>OSX</td>
<td>R</td>
<td>168.6</td>
<td>70.6</td>
</tr>
<tr>
<td>OSX</td>
<td>R+S600K</td>
<td>155.8</td>
<td>63.5</td>
</tr>
</tbody>
</table>

show that finetuning with SynBody outperforms that with AGORA across different baseline methods. The results highlight that SynBody demonstrates competitive traits as a training source.

Effectiveness of SMPL-XL. SMPL-XL enables layered human modeling which cannot be achieved with SURREAL/RenderPeople as Figure 5. We investigate the importance to synthesize data with these realistic considerations of SMPL-XL in Table 5. We compare SynBody with SURREAL-style data that renders body and cloth texture on human mesh surfaces, without actual cloth geometry or hair. It is observed that SURREAL-style data leads to significant performance degradation.

5.1.2 SMPL-X Estimation

Also known as expressive human pose and shape estimation, SMPL-X estimation requires recovery of body, hands, and face parameters.

Method and Training Data. In Table 6, we conduct experiments with the recent SoTA, OSX [28], as the base model. We compare the original baseline with training with a mixture of real datasets (“R” denotes COCO [29], MPII [3] and Human3.6M [16] with pseudo ground truth generated by NeuralAnnot [36]) and SynBody (“S600K” here denotes a downsampling of 600K instances as the real datasets are smaller for SMPL-X than that for SMPL estimation). Note that in the OSX paper, the values reported on 3DPW are obtained using a SMPL head. Here we standardize the output using the same SMPL-X head for both AGORA and 3DPW.

Main Results. We observe that adding SynBody in the training of OSX leads to a significant improvement in both overall estimation (more than 10 mm) and part-level (more than 5 mm for hands and face) estimation on AGORA. Moreover, in a fair comparison with the same model architecture, training with SynBody leads to 18 mm improvements in MPJPE on the 3DPW test set (here we follow the standard protocol to evaluate 14 key joints as 3DPW does not provide SMPL-X annotations). We speculate that it
Table 7: Benchmark of NeRF-based methods for 3D human neural rendering on SynBody.

<table>
<thead>
<tr>
<th>Method</th>
<th>Novel View</th>
<th></th>
<th>Novel Pose</th>
<th></th>
<th>Novel Identity</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR ↑</td>
<td>SSIM ↑</td>
<td>LPIPS ↓</td>
<td>PSNR ↑</td>
<td>LPIPS ↓</td>
<td>PSNR ↑</td>
</tr>
<tr>
<td>NeRF [34]</td>
<td>19.39</td>
<td>0.862</td>
<td>0.162</td>
<td>19.61</td>
<td>0.824</td>
<td>0.201</td>
</tr>
<tr>
<td>NeuralBody [41]</td>
<td>28.94</td>
<td>0.966</td>
<td>0.057</td>
<td>25.02</td>
<td>0.944</td>
<td>0.080</td>
</tr>
<tr>
<td>HumanNeRF [50]</td>
<td>28.32</td>
<td>0.963</td>
<td>0.066</td>
<td>21.97</td>
<td>0.879</td>
<td>0.108</td>
</tr>
<tr>
<td>AnimNeRF [40]</td>
<td>27.49</td>
<td>0.964</td>
<td>0.056</td>
<td>26.21</td>
<td>0.950</td>
<td>0.068</td>
</tr>
<tr>
<td>NHP [26]</td>
<td>25.66</td>
<td>0.953</td>
<td>0.076</td>
<td>24.18</td>
<td>0.945</td>
<td>0.080</td>
</tr>
</tbody>
</table>

is difficult to obtain accurate SMPL-X annotations, which is mitigated by SynBody as it provides large-scale high-quality SMPL-X labels paired with images that are rendered with diverse backgrounds and lighting conditions, which are beneficial to training a high-performing expressive human parametric model recovery.

5.2. Human NeRF

In this section, we benchmark popular NeRF-based methods for 3D humans on SynBody, validating the effectiveness and great potential of our dataset in human neural rendering. Our benchmark is built upon three perspectives according to the purpose of synthesis tasks, which can be categorized into novel view, novel pose, and novel identity.

Methods for Benchmark. We benchmark five methods in total, including the vanilla NeRF [34], NeuralBody [41] and HumanNeRF [50] for novel view synthesis, AnimNeRF [40] for novel pose synthesis, NHP [26] for generalizable human NeRF (novel identity synthesis). Except for NHP, all methods are trained in a person-specific manner, taking 4 views of the first 250 frames for training and the rest views and frames for evaluations.

Evaluation Protocols. We follow [34, 50] to evaluate all methods using three standard metrics: Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS [56]). To reach a consensus among all methods, all metrics are computed over the whole image with a black background.

Main Results. Benchmark results are reported in Table 7. We observe that all methods achieve comparable performances as real human data on SynBody. Models (NeuralBody and AnimNeRF) which rely on accurate SMPL estimation and blending weights perform better on novel poses. We attribute it to our layer-wise design which offers ground truth SMPL parameters for human NeRF training. Besides, we present visualization results in Figure 7, observing that the diverse motions and appearances, as well as loose garments in SynBody, pose further challenges in the field of neural rendering of 3D humans.

6. Conclusion

We present SynBody, a large-scale synthetic dataset that features a substantial number of subjects and high-quality 3D annotations. At the core is a clothed human model with multiple layers of representation. Our experiments demonstrate the effectiveness of SynBody on both human mesh recovery and human NeRF. Future research can leverage SynBody for developing and evaluating methods to predict body and cloth simultaneously. Furthermore, the high controllability of the synthetic dataset offers ample opportunities for further improvements, such as the incorporation of contact labels for human-scene interaction.

Figure 7: Novel view synthesis of different human NeRF methods on SynBody.
References


[34] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, and Michael J. Black. AMasa: Archive of motion capture as surface shapes. In ICCV, pages 5442–5451, 2019. 2, 6, 7


