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

Zhitao Yang[1,*] Zhongang Cai[1,2,*] Haiyi Mei[1,*] Shuai Liu[2,*] Zhaoxi Chen[3,*]
Weiye Xiao[1] Yukun Wei[1] Zhongfei Qing[1] Chen Wei[1]
Bo Dai[2] Wayne Wu[1,2] Chen Qian[1] Dahua Lin[2,4] Ziwei Liu[3,†] Lei Yang[1,2,†]
[1]SenseTime Research [2]Shanghai AI Laboratory
[3]S-Lab, Nanyang Technological University [4]The Chinese University of Hong Kong

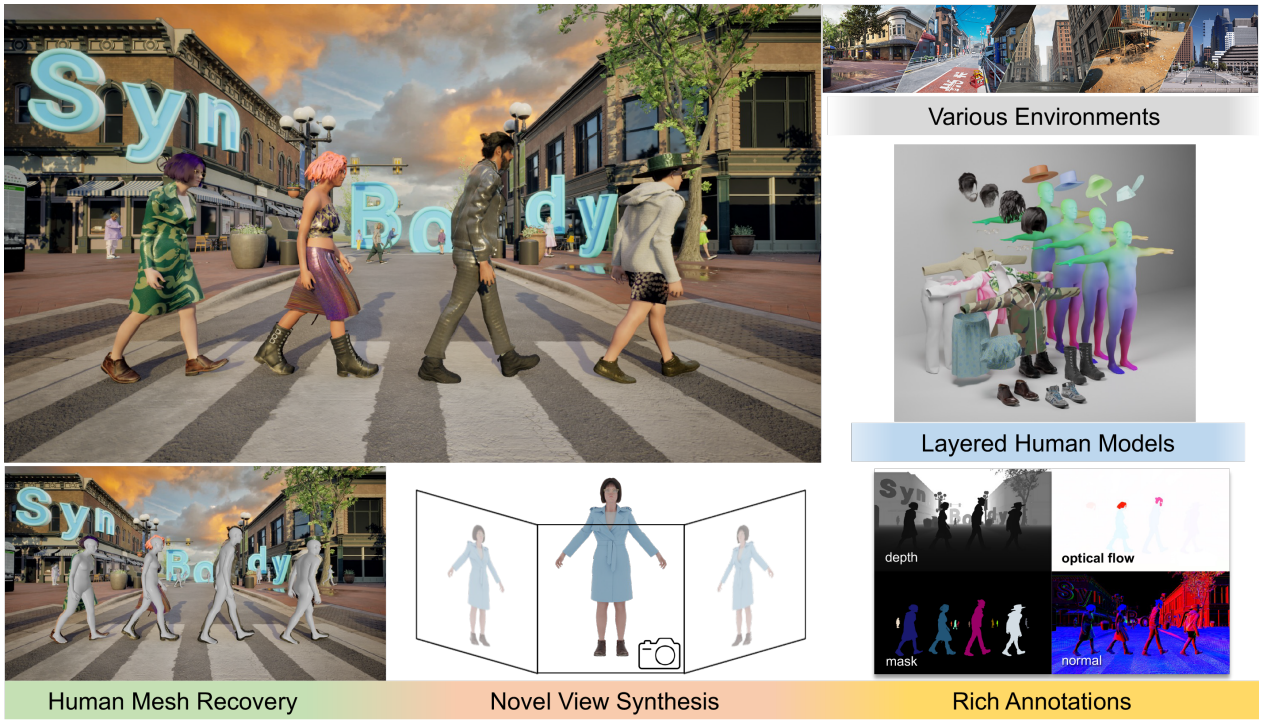

Figure 1: **SynBody is a large-scale synthetic dataset with massive number of subjects and high-quality annotations.** It supports various research topics, including human mesh recovery and novel view synthesis for human (Human NeRF).

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 1.7M images with corresponding accurate 3D annotations, covering 10,000 human body models, 1000 actions, and various viewpoints. The dataset includes two subsets for human mesh recovery as well as human neural rendering. Extensive experiments on SynBody indicate that it substantially enhances both SMPL and SMPL-X estima-*

*Equal contribution
†Corresponding author

tion. 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 [15–18,26,39,40] and human reconstruction [10, 19, 29, 30] 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 [3, 4,27,35] 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 shape, pose 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 1.7 million frames with corresponding ground-truth 3D human body annotations. It covers 10,000 human body models, 2,000 actions, 4 viewpoints, and 6 styles of scenes.

At the heart of SynBody is the layered parametric human model, which constructs the clothed human model in a bottom-up manner. SMPL-X [28] 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 LSMPL-X, a layered parametric human model based on SMPL-X. LSMPL-X enriches the SMPL-X model in three aspects: (1) Hair system: adding hair and beards to the FLAME [20] 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, LSMPL-X also adds rich textures for sampling various skin color and clothing texture.

The designed LSMPL-X is capable of 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 ensures that the annotations are naturally accurate, obviating the need for obtaining annotation via 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 [22]. 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 to images with corresponding annotations.

With SynBody, we launch two tracks that support human mesh recovery and human neural rendering, respectively. Experiments show that SynBody is more effective than AGORA under the same amount of training data for human mesh recovery. 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 [24]), 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 datasets 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 explicit cloth model, thus it provides layered 3D annotations of 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 the diversity and annotation quality for downstream tasks.

2. Related Works

Human Parametric Models. Several 3D human parametric models, such as SMPL [21], SMPL-X [28], and GHUM [38], have been developed to generate 3D human meshes from parameters that represent the human pose and shape using linear blend skinning. SMPL-X [28] extends SMPL [21] by combining FLAME [20] and MANO [31] 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.

[¶] https://maoxie.github.io/SynBody/

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.

Dataset	Type	ITW	Video	#Views	#SMPL	#Seq	#Subj.	#Motions	GT format
HumanEva [33]	R	-	✓	4/7	NA	7	4	6	3DJ
Human3.6M [11]	R	-	✓	4	312K	839	11	15	3DJ, SMPL
MPI-INF-3DHP [23]	M	✓	✓	14	96K	16	8	8	3DJ
3DPW [36]	R	✓	✓	1	32K	60	18	*	SMPL
Panoptic Studio [14]	R	-	✓	480	736K	480	∼100	*	3DJ
EFT [13]	R	✓	-	1	129K	NA	Many	NA	SMPL
ZJU-MoCap [30]	R	-	✓	21	180K	9	9	9	SMPL,mask
SURREAL [35]	S	✓	✓	1	6.5M	NA	145	2K	SMPL
AGORA [27]	S	✓	-	1	173K	NA	>350	NA	SMPL, SMPL-X, mask
HSPACE [3]	S	✓	✓	5	-	NA	100×16	100	GHUM/L, mask
GTA-Human [4]	S	✓	✓	1	1.4M	20K	>600	20K	SMPL
SynBody	S	✓	✓	4	2.7M	40K	10,000	2K	SMPL, SMPL-X, mask

Human Pose and Shape Estimation. Several methods [15–18, 26] have been proposed to estimate 3D human pose and shape parameters. HMR [15] directly regresses these parameters in an end-to-end manner, while SPIN uses an optimization step [18] to guide the learning process towards pseudo 3D labels. PARE [17] employs part attention to tackle occlusion, and VIBE [16] leverages temporal information in videos for SMPL estimation. To increase the data scale in a low-cost budget, several synthetic datasets have been proposed. SURREAL [35] renders textured SMPL body models in real-image backgrounds but does not account for cloth geometry, resulting in unrealistic subjects. AGORA [27] renders real human scans in a virtual world and provides high-quality synthetic data for image-based approaches. HSPACE [3] 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 [4] 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 [5, 7, 25, 32, 41] has been made to whole-body human pose and shape estimation. ExPose [5] introduces three experts to predict parameters for body, hand and face, and merge them in a copy-paste strategy. Hand4Whole [25] further improves the prediction of wrist and finger poses by leveraging selected hand joints 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 [27] provides image-based SMPL-X annotations by fitting SMPL-X on the scanned human models.

Human NeRF. NeRF [24] 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 [30] incorporates prior from a statistical body template to learn dynamic sequence, while Animatable NeRF [29] proposes to reconstruct an animatable human model that generalizes to new poses. Furthermore, NHP [19] leverages pixel-aligned features to generalize to unseen pose and subjects. Several datasets have been adapted to study human NeRF. ZJU-MoCap [30] captures 9 human subjects with 21 synchronized cameras, providing fitted human body model parameters as well as the foreground mask. H36M [12] collects 11 human subjects with 4 cameras, using marker-based motion capture system. A-NeRF [34] generates a synthetic dataset using SURREAL [35] to study factors that affect the 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 LSMPL-X 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, respectively, and features the same 10,000 diverse subjects.

3. Synthetic Data Generation System

As introduced in Figure 2, our infrastructure consists of 4 components: (1) a layered parametric human model creation service, a scalable process to generate layered human models, (2) a motion retargeting module to apply motions

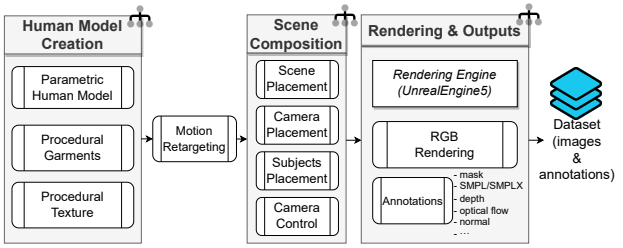

Figure 2: **Synthetic data generation system.** It consists of 4 components: 1) human model creation to generate layered human models, 2) motion retargeting to drive human models, 3) scene composition to place actors and cameras, and 4) rendering and outputs to generate multi-modal dataset.

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. By using our infrastructure, we can generate high-quality synthetic data for various computer vision tasks.

3.1. Layered Parametric Human Model

Parametric human models like SMPL-X [28] provide the ability of 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 [28] body model has a 3D mesh whose vertex locations are controlled by parameters for pose θ, shape β, and facial expression ψ. By modifying shape β, we obtain 3D meshes of various human heights and weights. And by alternating pose θ, 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 styles. Notice that different parts of garment pieces are connected with sewing lines, e.g., red line in Figure 3 (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 3 (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

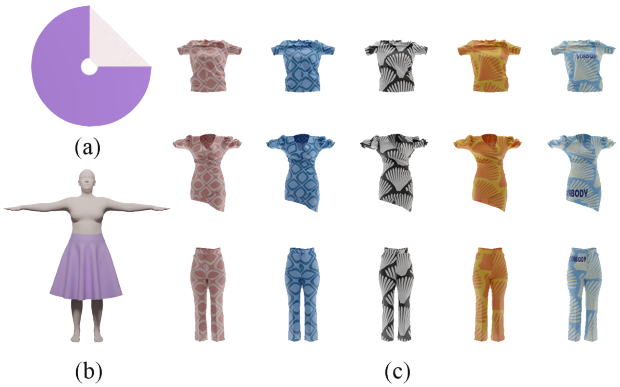

Figure 3: Demonstrations of Garment and Procedural Texture. (a) garment pattern with sewing lines, (b) simulated garment under the body in canonical space, (c) procedural textured garment whose style and color are elevated by layered masks provided by patterns and decals.

particle system to generate realistic hair on a head-shaped mesh. To achieve this, a template of hairstyle or facial hair T_{hair} would attach to the vertices on the mesh which are marked with different areas V_{hair} (including fringe, top, temporal bone, occipital bone, and bottom area). Designers draw multiple sets of guidelines L_{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 P_{length} and curliness $P_{curliness}$. The entire hair system is composed of multiple sub-particle systems $\{V_{hair}, L_{guide}, P_{length}, P_{curliness}\}$.

Accessories. We also add template accessories $T_{accessories}$ 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 β 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 textures $T_{procedural} = \{T_{pattern}, T_{decals}, T_{bump}, P_{mapping}\}$ used in clothing are composed of multiple pre-set textures by alpha blending as demonstrated in Figure 3 (c). The pattern texture $T_{pattern}$ and decals texture T_{decals} serve as layered masks to elevate the style and color, while the bump texture T_{bump} functions as a height map, adding detailed normal information to the clothing's surface. The mapping parameters $P_{mapping}$ control the coordination of all the textures in the UV space. Besides, we build a similar texture template for SMPL-X, in which we pre-draw layered masks to separate different body features, such as skin, lips, eyebrows, and eyes, allow-

SURREAL [35] RenderPeople [2] LSMPL-X

Figure 4: A Demonstration of comparison between commonly used human models in existing synthetic datasets [3, 4,27,35] and our LSMPL-X. We obtain high-quality models equal to RenderPeople [2], both are much better than SURREAL [35], and our model has the capability of scaling up easily by random various assets and body shapes.

ing for color adjustment and blending in specific regions.

Combining all the elements above, as shown in Figure 4, 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 various sources, such as academic motion datasets, motion captures, and artist-crafted sources, to LSMPL-X model skeletons. Despite variations in bone names, bone lengths, and rest pose bone rotations, the source skeletons are typically structurally similar to the target skeletons.

Pose frames in motion clips contain root translation $T(t)$ and rotations of each bone in the corresponding parent bone's space $\{R_0(t), R_1(t), ..., R_n(t)\}$. Following the forward kinematics (FK) manner, each bone's rotation relative to the model space can be calculated:

$$\widehat{R}_i(t) = \widehat{R}_{p(i)}(t) \cdot R_i(t) \tag{1}$$

where $\widehat{R}_i(t)$ is the rotation of $R_i(t)$ in model space at t frame, and $p(i)$ indicates the parent of i. Specially, root bones have no parents so $\widehat{R}_0(t) = R_0(t)$.

To retarget motion from one skeleton to another, we assume the source and the target motion can drive corresponding bones of both skeletons to the same rotation in model space. We have T-pose frames of skeletons by manually posing them as T-posing, and treat T-pose as all motions' first frame pose. So motions relative to their T-pose frame can be easily obtained:

$$\widehat{R}_i(t) = \widehat{R}_i(0) \cdot \widehat{R}_{i\_src}(t) \cdot \widehat{R}^{-1}_{i\_src}(0). \tag{2}$$

Considering skeletons have different bone lengths, which could result in "sliding feet" artifacts on target skeletons. We simply scale T according to the ratio of pelvis bones' height to mitigate them: $T(t) = \frac{H_{pelvis}}{H_{pelvis\_src}} \cdot T_{src}(t)$.

SMPL-X Annotations. Considering LSMPL-X body shapes are sampled frome SMPL-X, the shape β can be derived directly from the corresponding SMPL-X. Pose θ in the model space is calculated in the motion retargeting module. And in the scene composition module, models are placed in a 3D scene with world space locations T_w and rotations R_w. Pose θ_w in world space is calculated by applying those world transformations to θ. So SMPL-X annotations are constructed with $\{\beta, \theta_w\}$.

SMPL Annotations. Although LSMPL-X naturally provide accurate SMPL-X annotations, SMPL cannot be derived directly. Thus, we need to refit the SMPL parameters. The optimization process consists of two steps. First, we fit shape β of all human models under the T-pose to its corresponding SMPL-X. Secondly, for each sequence, it is initialized with fitted β and its original pose. we fix β while fitting the body pose θ for SMPL. More details can be found in the Sup. Mat.

3.3. Scene Placement

To place N_c 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 $p_i^* = \{x_i^*, y_i^*\}$ of a character with the shape of $\{l_i, w_i\}$ should satisfy:

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

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

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_a 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_{max} is set to prevent an unreasonably small proportion of subjects in the image. To control the visibility of subjects, L_{min} is defined as $\frac{\lambda}{sin(\alpha/2)} \underset{i}{max} ||p_i - \bar{p}||_2$, where $i = 1, ..., N_c$, α is the field of view of the camera, and λ 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.

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 [9], 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 LSMPL-X models randomly created with different body shapes and genders. Each model is then combined with the following assets: (1) hairstyles T_{hair} sampled from 45 particle hairs; (2) garments G_{tmp} sampled from 68 clothing models containing multiple outfits; (3) procedural texture $T_{procedural}$ generated by sampling $T_{pattern}$, T_{decals}, and T_{bump} from 1,038 template textures, and color values are randomly sampled; (4) accessories $T_{accessories}$ sampled from 46 template assets. For each sequence, 1 to 4 models are randomly selected and applied with 2 to 10 seconds of continuous motion in SMPL-X format, obtained from AMASS [22]. The system generates 40,000 sequences and 1.7M images with annotations, utilizing 6 vast and realistic scenes created by professional artists. 2.7M SMPL/SMPL-X annotations are provided, excluding highly occluded subjects.

4.2. Human Mesh Recovery

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 [6], we generate bounding boxes from the resulting 2D keypoints. Particularly, for SMPL-X training, we use occlusion analysis (see Sec. 3.4) to filter out over-occluded bounding boxes for hands.

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 1024×1024, where the motion sequence is randomly sampled from AMASS [22]. 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

5.1. SMPL Estimation

Body Model and Baselines. The experiments in this study employ the SMPL model [21], a differentiable function that represents humans as 3D triangle meshes denoted by $M(\theta,\ \beta)\ \in \mathbb{R}^{6890\times 3}$. This model representation comprises of pose parameters denoted by $\theta \in \mathbb{R}^{72}$ and shape parameters denoted by $\beta \in \mathbb{R}^{10}$. Our study focuses on training and evaluation of three image-based methods of HMR [15], SPIN [18], and PARE [17] and one video-based method of VIBE [16].

Evaluation Metrics. For evaluating 3D human pose estimation, this study employs two commonly used metrics, namely **MPJPE** (Mean Per Joint Position Error) and **PA-MPJPE** (Procrustes Aligned Mean Per Joint Position Error). The MPJPE is calculated as L_2 distance averaged over all joints, while PA-MPJPE aligns the predicted keypoints to match the ground-truth with the Procrustes method [8], before computing the MPJPE between the aligned pose and ground truth. Additionally, we also calculate **PVE** (Per Vertex Error), which is the average distance between predicted and ground-truth mesh vertices.

Results on 3DPW testset. Table 2 illustrates three main observations: **(1) Training strategy.** We experiment with two training strategies, namely Blended Training (BT) and Fine-tuning (FT). The former train synthetic and real data jointly,

Table 2: Training results on popular baselines with real and SynBody datasets on 3DPW Testset. "R" means the public academic datasets of real human and "S" means our SynBody data.

Method	BT/FT	Datasets	MPJPE	PA-MPJPE	PVE
HMR [15]	-	R	112.30	67.50	141.92
HMR	BT	R + S	102.67	60.81	119.02
HMR	FT	R + S	**95.01**	**57.62**	**116.10**
SPIN [18]	-	R	96.90	59.20	119.70
SPIN	FT	R + S	**84.14**	**53.67**	**103.79**
PARE [17]	-	R	81.79	49.36	105.27
PARE	FT	R + S	**78.98**	**48.46**	**103.86**
VIBE [16]	-	R	94.88	57.08	108.59
VIBE	BT	R + S	**93.04**	**57.00**	**107.23**

while the latter forms the training in a two-stage procedure, which first trains with real data and then finetune with a combination of real and synthetic data. Table 2 demonstrates that no matter FT or BT applies, significant performance gain is obtained compared to training with only real data; **(2) Baseline Model.** Consistent improvements are observed across different baseline models, even when the recent method [17] achieves low errors, it still obtains 2.81% and 0.9% on MPJPE and PA-MPJPE, respectively. **(3) Video-based Method.** As SynBody naturally consists of videos, it can also be used for training video-based approaches. Experiments on VIBE show improved results on all metrics when compared to the baseline.

Table 3: Training results on popular baselines with real and SynBody datasets on AGORA validation set [27]. "R" means the public academic datasets of real human and "S" means our SynBody data.

Method	Datasets	MPJPE	PA-MPJPE	PVE
HMR	R	226.71	87.72	248.35
HMR	R + S	**199.51**	**77.97**	**210.37**
SPIN	R	212.91	79.76	217.88
SPIN	R + S	**196.81**	**76.06**	**205.83**
PARE	R	178.15	67.13	189.73
PARE	R + S	**169.93**	**64.37**	**179.81**

Results on AGORA validation set. AGORA [27], a recent synthetic dataset, has draw more attention as its more challenging than 3DPW in terms of environmental and person-person occlusion. Table 2 illustrates that training with SynBody exhibits remarkable improvements across methods, when compared with the baseline models in AGORA validation set.

The effect of data scale. As our generation system can easily scale the data size while maintaining data diversity, we conduct blended training with HMR to study the influence of synthetic data scale. Figure 5 demonstrates that adding more SynBody data generally leads to better performance. These experiments confirm that synthetic data is a valuable complement to real data and serves as a readily scalable training source to supplement the typically limited real data.

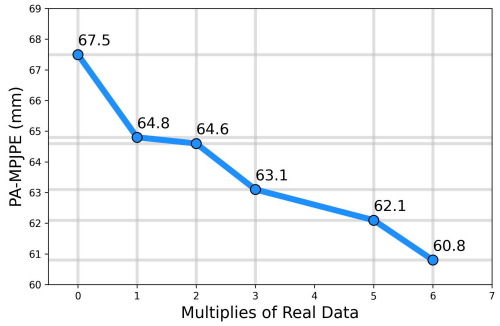

Figure 5: Impact of SynBody Data. The horizontal axis represents the amount of SynBody data is varied by multiples of real data. Baseline model is HMR.

Table 4: Comparison between SynBody-10W and AGORA on 3DPW test set. "A" means AGORA, and "S10W" means SynBody-10W which is a subset of SynBody datasets with over 10W SMPL annotations. For fair comparison, the total number of SMPL annotations is exactly the same as that of AGORA.

Method	Datasets	MPJPE	PA-MPJPE	PVE
HMR	R + A	101.68	57.43	124.16
HMR	R + S10W	**95.10**	**57.20**	**117.86**
SPIN	R + A	88.44	54.97	110.35
SPIN	R + S10W	**84.32**	**53.63**	**104.20**
PARE	R + A	85.34	48.39	109.77
PARE	R + S10W	**80.42**	**47.68**	**104.08**

Comparison with AGORA training set. To demonstrate the efficacy of data diversity and high-quality annotations, we sample a subset, "SynBody-10W", comparable in size to AGORA. Results presented in Table 4 indicate that training with SynBody outperforms training with AGORA using three different methods. The findings emphasize the significance of synthetic data's diversity and annotation quality for achieving superior performance with equivalent data volume.

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 5 methods in

Table 5: Benchmark of NeRF-based methods for 3D human neural rendering on SynBody.

Method	Novel View			Novel Pose			Novel Identity		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [24]	19.39	0.862	0.162	19.61	0.824	0.201	-	-	-
NeuralBody [30]	28.94	0.966	0.057	25.02	0.944	0.080	-	-	-
HumanNeRF [37]	28.32	0.963	0.066	21.97	0.879	0.108	-	-	-
AnimNeRF [29]	27.49	0.964	0.056	26.21	0.950	0.068	-	-	-
NHP [19]	25.66	0.953	0.076	24.18	0.945	0.080	22.46	0.927	0.103

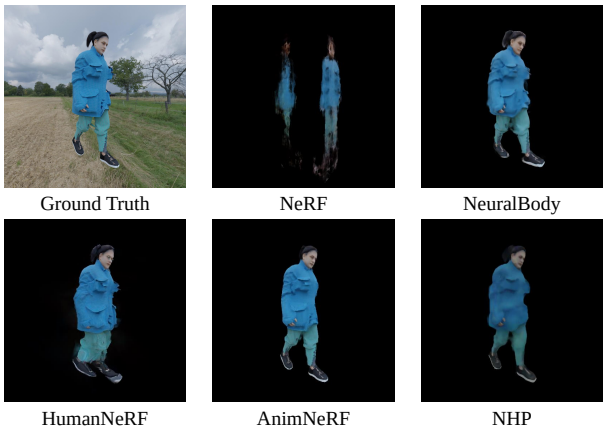

Figure 6: Novel view synthesis of different human NeRF methods on SynBody.

total, including the vanilla NeRF [24], NeuralBody [30] and HumanNeRF [37] for novel view synthesis, AnimNeRF [29] for novel pose synthesis, NHP [19] for generalizable human NeRF (novel identity synthesis). Except 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 [24, 37] 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 [42]). 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 5. 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 6, 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 synthetic dataset offers ample opportunities for further improvement, such as the incorporation of contact labels for human-scene interaction.

Societal Impacts. Although SynBody is a synthetic dataset, the assets used for its generation may be unbalanced. While hairstyles and skin color are sampled randomly, they do not result in racial bias. However, the remaining assets, such as body shape and clothing, may be imbalanced, leading to potential bias in the generated human models.

References

[1] Marvelous designer, 2023. https://www.marvelousdesigner.com. 4

[2] Render people, 2023. https://renderpeople.com. 5

[3] Eduard Gabriel Bazavan, Andrei Zanfir, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Hspace: Synthetic parametric humans animated in complex environments. *arXiv preprint arXiv:2112.12867*, 2021. 2, 3, 5

[4] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, and Ziwei Liu. Playing for 3d human recovery. *arXiv preprint arXiv:2110.07588*, 2021. 2, 3, 5

[5] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 20–40. Springer, 2020. 3

[6] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. 24:663–676, 2019. 6

[7] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of

expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, pages 792–804. IEEE, 2021. 3

[8] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 1975. 6

[9] Shawn Hargreaves and Mark Harris. Deferred shading. In *Game Developers Conference*, volume 2, page 31, 2004. 6

[10] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 2

[11] 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 transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 3

[12] 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 Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 3

[13] H. Joo, N. Neverova, and A. Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. *ArXiv*, abs/2004.03686, 2020. 3

[14] H. Joo, Tomas Simon, Xulong Li, H. Liu, L. Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart C. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:190–204, 2019. 3

[15] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2, 3, 6, 7

[16] 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. 2, 3, 6, 7

[17] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. *arXiv preprint arXiv:2104.08527*, 2021. 2, 3, 6, 7

[18] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 2, 3, 6, 7

[19] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3, 8

[20] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2

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

[22] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 2, 6

[23] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 3

[24] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3, 8

[25] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2308–2317, 2022. 3

[26] Hui En Pang, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 3

[27] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2, 3, 5, 7

[28] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2, 4

[29] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 2, 3, 8

[30] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2, 3, 8

[31] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2

[32] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1749–1759, 2021. 3

[33] L. Sigal, A. O. Balan, and Michael J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87:4–27, 2009. 3

[34] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Advances in Neural Information Processing Systems*, 2021. 3

[35] Gül Varol, J. Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, 2017. 2, 3, 5

[36] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 3

[37] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, June 2022. 8

[38] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 2

[39] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. Deciwatch: A simple baseline for 10× efficient 2d and 3d pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 607–624. Springer, 2022. 2

[40] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: a plug-and-play network for refining human poses in videos. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 625–642. Springer, 2022. 2

[41] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *arXiv preprint arXiv:2207.06400*, 2022. 3

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 8