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

Zhitao Yang[1,*] Zhongang Cai[1,2,3,*] 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[2]
Chen Qian[1] Dahua Lin[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
[*]Equal Contribution [†]Corresponding Author
https://synbody.github.io/

A. Details of Synthetic Data Generation

A.1. Construction of AMASS subset

To ensure the validity of the motions, we selected a subset from AMASS [1]. Following the BABEL annotations [2], we excluded interactive motions, non-ground motions, and motions with a duration of less than 2 seconds. The specific categories excluded were: "unknown", "interact with/use object", "touching body part", "exercise/training", "move up/down incline", "sit", "touch object", "touching face", "swim" and 'fall.' The resulting subset comprised 1,187 motion sequences, each lasting more than 2 seconds.

A.2. Details of Camera Placemeent

Figure 1 illustrates the process of calculating the camera's minimum distance to the subjects' center. The shaded box represents the envelop box of a subject across frames, while the gray circle ensures that it encompasses all subjects. Therefore, the camera should ensure that the circle is within its field of view, which requires the distance to be greater than L_{min}, the radius of the red circle. As stated in the main text, $L_{min} = \frac{\lambda}{sin(\alpha/2)} \underset{i}{max} ||pv_i - \bar{p}||_2$. In this equation, pv represents the position of the vertex of the envelope box of the subject, with $i = 1, ..., N_v$, where N_v is the total number of vertices. Additionally, cameras that have a pitch angle outside of a predefined range which is $[-5°, 30°]$ will be excluded from consideration, L_{max} is set to 10 meters for preventing an unreasonably small proportion of subjects in the image.

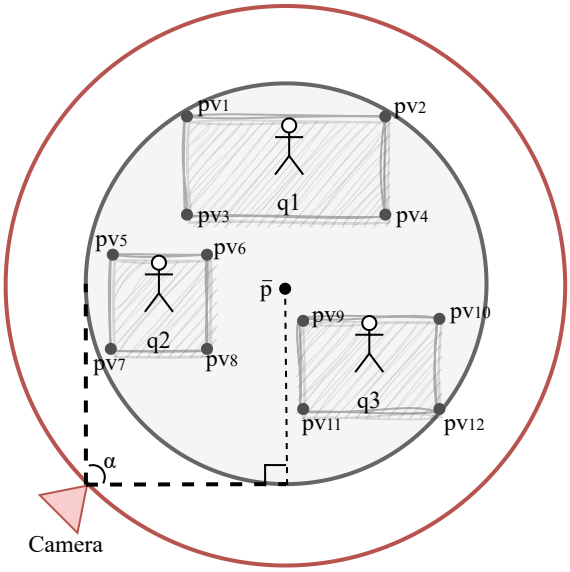

Figure 1: Illustration of camera placement.

B. Data examples of SynBody

B.1. Image examples

In the main paper, we show 1.2M images (2.7M instances in 27K sequences) containing subjects of neutral gender. However, SynBody also contains rendered images with three genders (neutral, female, male) that add up to ∼1.6M images (6M instances in 38K sequences) in the grand total. We show image examples in Figure 2.

B.2. Annotation examples

SynBody features accurate and diverse annotations that support various human perception and reconstruction tasks. In Figure 3, we show an RGB image with paired labels such

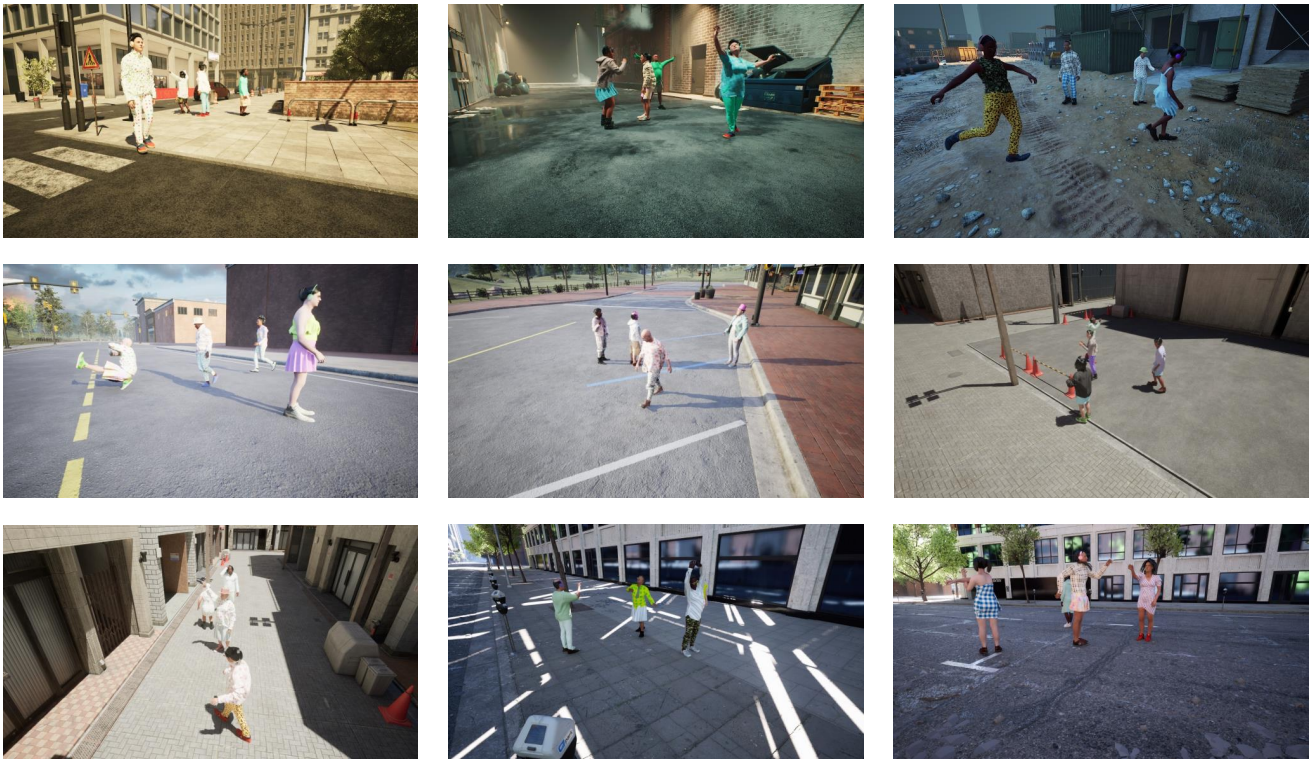

Figure 2: Illustration of synthetic images. SynBody features subjects with a variety of appearances and poses. These subjects are captured from various camera angles, set against diverse, realistic backgrounds, and illuminated under different lighting conditions. These considerations are critical to the usefulness of SynBody across various tasks.

as segmentation masks, keypoints, normal map, and SMPL-X. We highlight that some labels are expensive to obtain in real life, making SynBody a promising alternative for scaling up training data.

C. Asset examples of SynBody

SynBody utilizes a wide range of 3D assets in the rendering. These assets enhance the realism and diversity of generated images.

C.1. Scenes

In Figure 4, we place our virtual subjects in expansive, meticulously crafted 3D scenes. These environments are not only vast in scale but also emulate lifelike atmospheres, capturing a myriad of architectural designs from various cultures. We argue that the intrinsic diversity and high-fidelity quality of these backgrounds not only enhance the visual appeal but also play a pivotal role in potentially mitigating the synthetic-real domain gap.

C.2. Hairstyles, Clothes, and Accessories

One of the standout features of SMPL-XL is the extensive collection of appearance elements beyond the naked human body mesh. In Figure 5, we demonstrate a vast repository of diverse hairstyles, clothing (with procedural textures, and accessories (such as glasses, shoes, hats, and headphones). These elements enhance the depth of detail and customization available in our SMPL-XL: a comprehensive layered human representation.

References

[1] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019. 1

[2] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *CVPR*, pages 722–731, 2021. 1

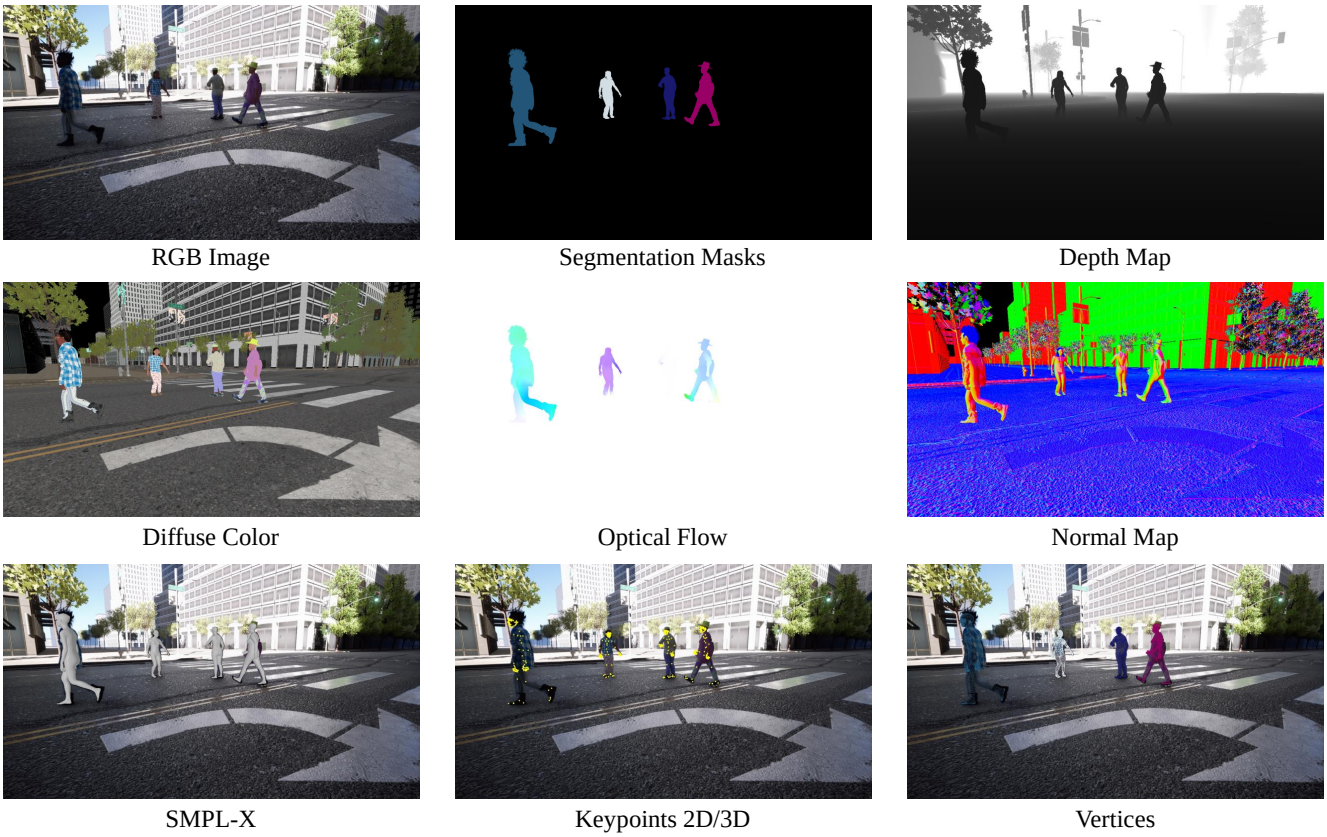

Figure 3: Illustration of annotations. SynBody provides accurate annotations paired with RGB images. Therefore, SynBody can support a myriad of human-related tasks in perception and reconstruction.

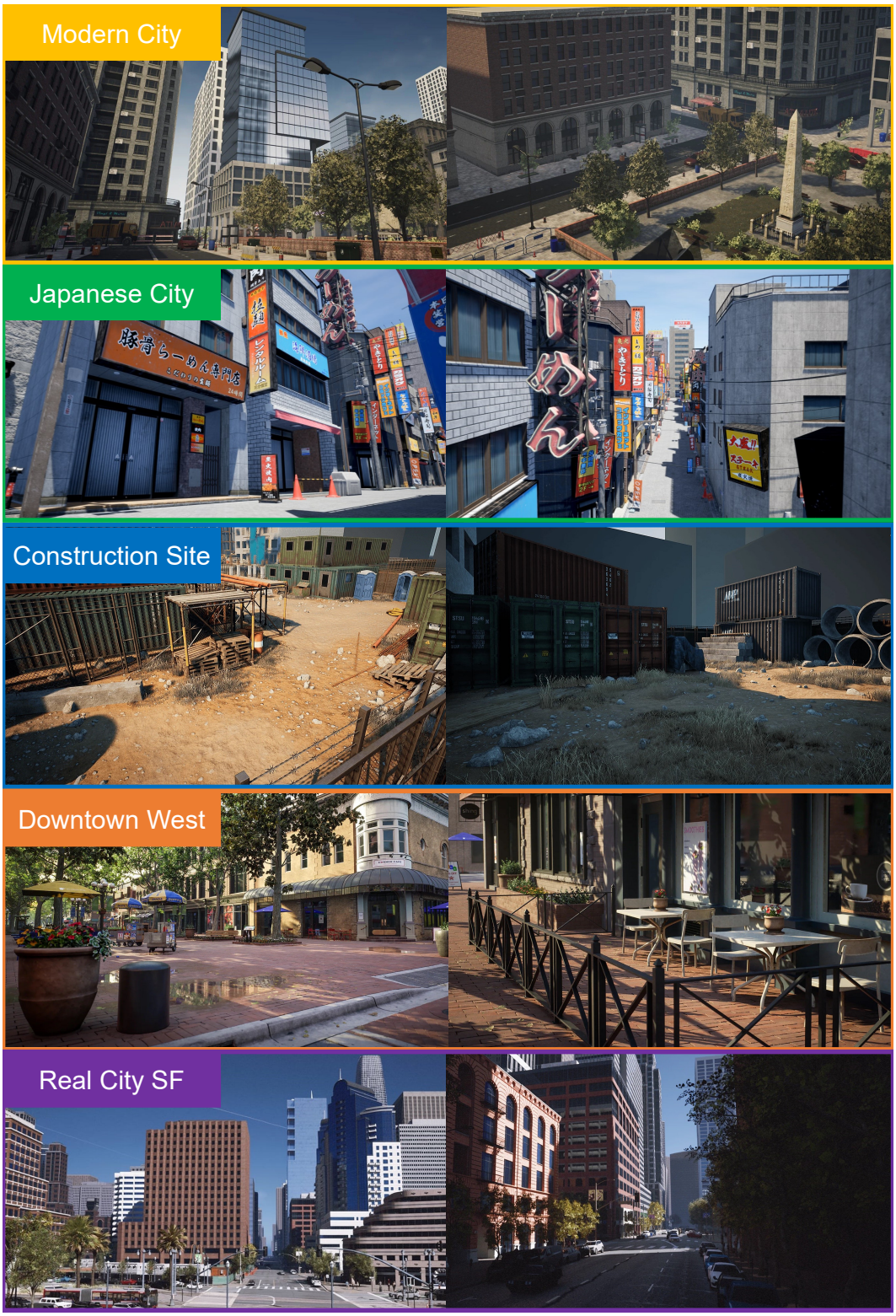

Figure 4: Illustration of scenes. We utilize high-quality, diverse city-scale scene models in rendering our images.

Figure 5: Illustration of Assets used in SMPL-XL. SMPL-XL enables a layered human modeling that encompasses a wide range of hairstyles, accessories (such as hats and shoes), and clothes of different types, dimensions, and textures.