SCULPT: Shape-Conditioned Unpaired Learning of Pose-dependent Clothed and Textured Human Meshes

Soubhik Sanyal[1] Partha Ghosh[1] Jinlong Yang[1*]
Michael J. Black[1] Justus Thies[1,2] Timo Bolkart[1*]
[1]Max Planck Institute for Intelligent Systems [2]Technical University of Darmstadt
{soubhik.sanyal, partha.ghosh, jinlong.yang, black, justus.thies, timo.bolkart}@tue.mpg.de

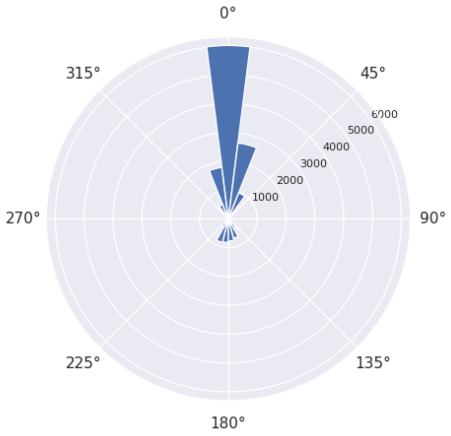

Figure 1. Histogram of the body rotations of the used training corpus with respect to the camera view (0° is frontal).

1. Dataset statistics:

Refer to Fig. 1 for the view statistics of the *SCULPT dataset*. It is observed that the dataset exhibits a bias towards both frontal and near-frontal views. The dataset offers a variety of clothing types, including 'short sleeve T-shirt/short trouser", "short sleeve T-shirt/long trouser", "long sleeve T-shirt/long trouser", "long sleeve T-shirt/short trouser", "shirt/long trouser", and "shirt/short trouser" which are similar to the assortment found in the CAPE dataset. The labels were automatically generated using CLIP, as detailed in Sec. 3.2. The dataset contains 2,483 "short-short", 6,260 "short-long", 335 "long-short", 3,425 "long-long", 939 "shirt-short", and 2,920 "shirt-long" items, where "short-short" refers to "short sleeve T-shirt/short trousers", and so forth. Regarding the color types in the training dataset of fashion images, there are descriptions for 115 different colors. Examples of these colors include red, blue, green, khaki, pink, peach, and tan, among others. We plan to release the dataset annotations for research purposes.

*Now at Google.

2. BLIP texture description accuracy:

In a perceptual study with 2000 labeled images on Amazon Mechanical Turk, BLIP labels were judged to be correct 92.7% of the time for upper body clothing and 89.7% for lower body clothing. During the study, participants received images alongside associated BLIP labels and assessed their validity with a "yes/no" answer. We treat the human judgements as ground truth. The common point of mismatch between the participants and BLIP labels was nearby colors like khaki or tan etc.

3. Parameters of the differentiable rendering:

Our model utilizes PyTorch3D's soft rasterizer for differentiable rendering, with a zero blur radius for one-to-one pixel-triangle correspondence. A directional light with fixed intensity and orthographic projection is employed for mesh rendering. Body orientation or view for each rendered image is randomly chosen from the training dataset, with this randomization applied per image in each batch during generator forward passes. Check our training and inference codebase for reproducibility.