SiCloPe: Silhouette-Based Clothed People – Supplementary Materials

Ryota Natsume[1,3*] Shunsuke Saito[1,2*] Zeng Huang[1,2] Weikai Chen[1]
Chongyang Ma[4] Hao Li[1,2,5] Shigeo Morishima[3]

[1]USC Institute for Creative Technologies [2]University of Southern California
[3]Waseda University [4]Snap Inc. [5]Pinscreen

1. Implementation Details

1.1. Deep Visual Hull

We employ a network that is similar to [1] for our proposed deep visual hull algorithm. The detailed network structure is shown in Figure 1. The network consists of two parts: (1) a convolutional neural network for extracting features from a given view and 3D query points, and (2) a classification network that consumes the multi-view features and predicts per-point probability of lying inside the reconstructed object. The input to our network includes images from 12 different views. For each view, the input is a four-channel image, which is the concatenation of the previously synthesized silhouette mask (one channel) and the 2D pose map at this view (color-coded in three channels).

The feature extraction network is composed of four convolutional layers with a kernel size of 5×5 and channel sizes of 4, 8, 16, 32, as well as two fully connected layers with a dimension of 128 and 256 for the hidden layer, respectively. After each convolutional layer, we also apply ReLU activation. For each 3D query point, we first project it onto the image plane of each view and extract the features at the projected location from the output of each layer as well as the input layer. The extracted features are concatenated and passed to the fully connected layers, resulting a 256-dimension feature vector for each view (Figure 1(a)).

As shown in Figure 1(b), the classification network first concatenates feature vectors from all input views and then applies max pooling to obtain a view-independent latent code of length 256. Finally, the latent code is fed into a four-layer MLP network for inferring the per-point probability of staying inside the object surface.

1.2. Baseline Methods

To validate our design choice, we compare our silhouette-based reconstruction with volumetric reconstruction using voxels [4]. Additionally, we evaluate our input of silhouette by comparing with results from RGB input. We describe the implementation details of these baseline methods below.

For 2D silhouette synthesis using RGB input, we use a network architecture based on U-Net [2] by replacing the original single-channel segmentation with RGB images in our proposed network. We use the same loss function and optimizer as our silhouette synthesis network. The voxel prediction network is based on a stacked hourglass network [3]. This network takes as input silhouette/image ($\{1, 3\} \times 256 \times 256$), 2D pose ($3 \times 256 \times 256$), and 3D pose ($304 \times 64 \times 64$), where the joint heat maps in depth for each joint are concatenated into the channel dimension ($16 \times 19 = 304$). Here we use two stacks for both silhouette-input and RGB-input cases. Following [5], we concatenate the 3D pose information after a 4x downsampling operation by pooling in the network. The network predicts an occupancy field of human body of resolution $64 \times 64 \times 64$, which is optimized using a BCE loss $\mathcal{L}_{vol}$ between the ground truth and the prediction together with an additional reprojection loss from the front view $\mathcal{L}_{pf}$ and the side view $\mathcal{L}_{ps}$ (see Figure 2). The reprojection loss computes BCE loss between the ground truth silhouettes and 2D projected voxels along x and z axis using $\max$ operation, constraining the resulting silhouettes from each view to be consistent with ground truth [4]. The total loss function is given by

$$\mathcal{L} = \mathcal{L}_{vol} + \lambda_p \cdot (\mathcal{L}_{pf} + \mathcal{L}_{ps}),$$

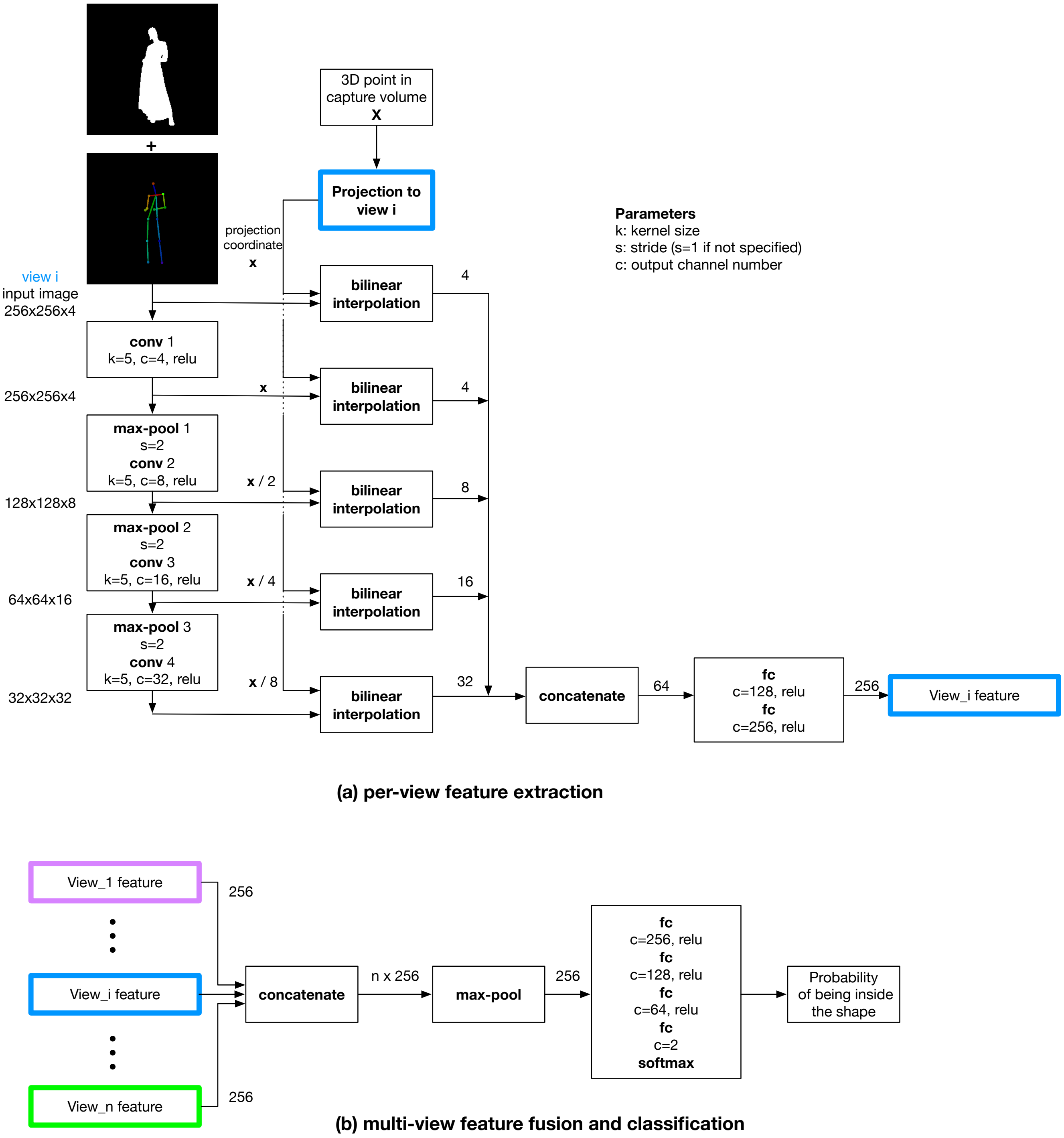
where the relative weight λ_p is set to 0.1 in our experiments. We use RMSProp optimizer with a learning rate of 2.0×10^{-4} and a batch size of 4. Note that this ablation study uses only frontal views as input for simplicity.

2. Additional Evaluations

2.1. Comparison with Voxel Representation

Figure 3 shows qualitative comparisons between our silhouette based shape representation and an alternative voxel based representation. Note that ours is trained using silhouettes and 2D poses as input, while the voxel regression uses 2D silhouettes, 3D poses, and 2D poses as input.

[*] Joint first authors



(a) per-view feature extraction

(b) multi-view feature fusion and classification

Figure 1: Our network architecture for the deep visual hull computation.

Compared to direct predicting the occupancy of 3D voxels, our implicit shape representation based on 2D silhouettes leads to more faithful reconstruction results with much smaller errors.

2.2. Ablation Study on Silhouette Synthesis

Figure 4 and Table 1 show the ablation study on the design choice of our silhouette synthesis network. To validate the importance of 2D pose information of the input view, we train the same silhouette synthesis network using the same

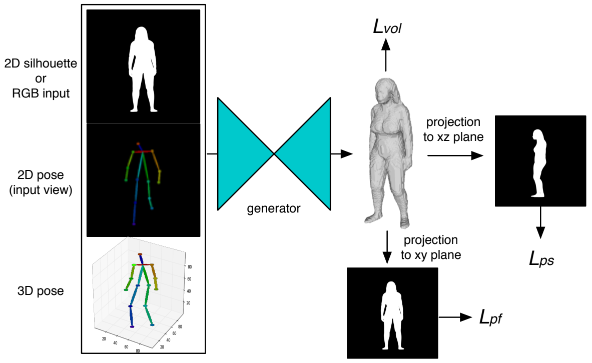


Figure 2: Illustration of the baseline voxel regression network.

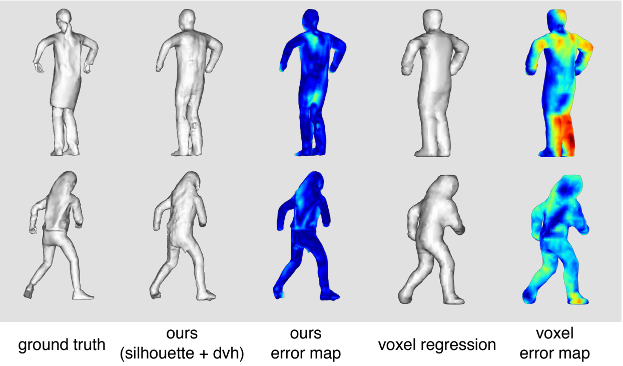


Figure 3: Qualitative evaluation of our silhouette-based shape representation as compared to direct voxel prediction.

training dataset	output	w/ input view pose	IoU (2D)
RenderPeople	single view	yes	**0.882**
RenderPeople	single view	no	0.875
RenderPeople	all views	yes	0.806
SURREAL	single view	yes	0.782

Table 1: Ablation study of our silhouette-based representation.

configuration but without the 2D pose of the input view. The reconstruction accuracy is evaluated by computing mean Intersection over Union (IoU) using the subjects in our test set from the predefined 12 views spanning every 30 degrees in yaw axis.

The model without 2D pose from the input view has difficulty associating loose clothes (e.g., dresses) with the novel view points, impairing the overall performance (see the fourth column in Figure 4).

We also train a silhouette synthesis network that predicts a set of silhouettes from a set of predefined view points in one go, instead of independently predicting silhouettes from each view point together with 2D joint information from the target view. The network generates the silhouettes from the predefined 12 view points at once. All the other configurations are identical to our main algorithm. This alternative approach also fails to produce plausible silhouettes and severely overfits to the training data samples (see the fifth column of Figure 4).

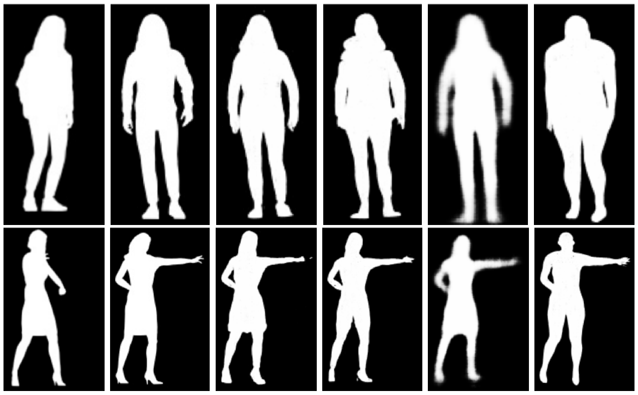


Figure 4: Qualitative evaluation of different silhouette synthesis methods. From left to right: silhouette of the input view, ground-truth silhouette of the target view, results of our full algorithm, results without 2D pose information, results from a set of predefined view points, and the ones by training on the SURREAL dataset [5].

Lastly, we demonstrate the importance of our clothed human training dataset to faithfully capture subjects with various clothes. We train our proposed network on the SURREAL dataset [5] in which all the subjects are in tightly-fitting clothes. We randomly select $14,490$ meshes from the training set of SURREAL and train our silhouette synthesis network with the same configurations as ours. Due to the lack of various cloth details, the resulted model is unable to predict plausible silhouettes with loose clothes (see the last column of Figure 4).

3. Additional Results

Figure 5 shows our 3D reconstruction results of clothed human body using test images from synthetic rendered data. Those test images have not been used for training. For each image, we show the back-view synthesis result, the reconstructed 3D geometry with with plain shading, as well as the fully textured output mesh rendered from a different view point.

4. Example Training Data

Figure 6 shows a collection of our rendered examples with both frontal and back views. Our networks are trained based on these synthetic data and can generalize well to handle real test images.

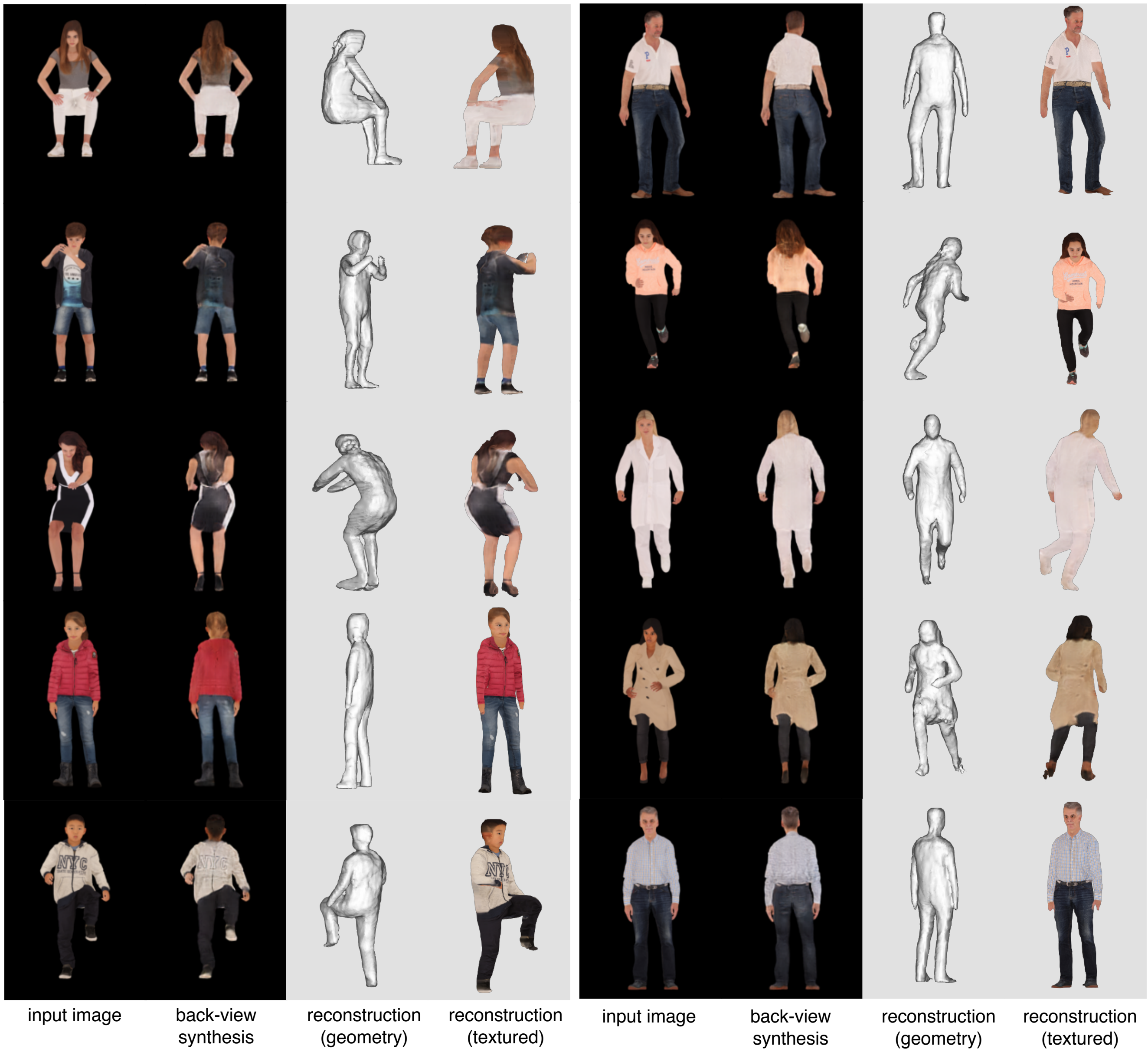

input image
back-view synthesis
reconstruction (geometry)
reconstruction (textured)
input image
back-view synthesis
reconstruction (geometry)
reconstruction (textured)

Figure 5: Our 3D reconstruction results of clothed human body using test images from the synthetically rendered data.

References

[1] Z. Huang, T. Li, W. Chen, Y. Zhao, J. Xing, C. LeGendre, L. Luo, C. Ma, and H. Li. Deep volumetric video from very sparse multi-view performance capture. In *European Conference on Computer Vision*, pages 336–354, 2018.

[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[3] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499, 2016.

[4] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision*, pages 20–36, 2018.

[5] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017.

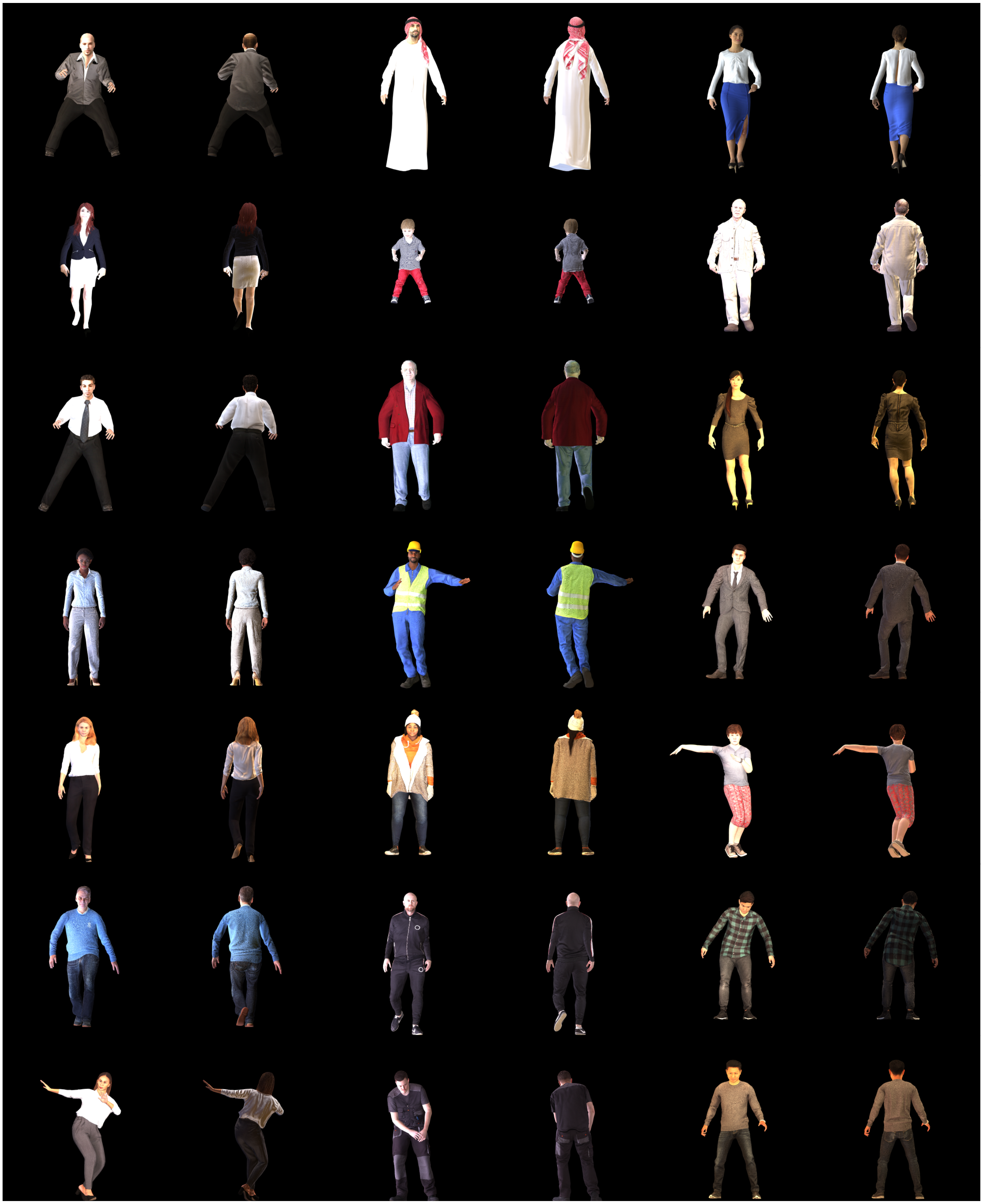


Figure 6: Our synthetically rendered training samples in our dataset.