Temporal Shape Transfer Network for 3D Human Motion (Supplementary Material)

João Regateiro Edmond Boyer
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France
name.surname@inria.fr

1. Introduction

In this supplementary material, we detail the network architecture (Figure 1 and Table 1) and present additional qualitative results (Figures 6, 7 and 8).

The proposed network is trained and tested on 18 and 9 randomly generated body shapes, Figures 2 and 3 represent the synthetic shapes used for training and testing, respectively. To further validate the generalization ability to unseen shapes we use realistic body models from FAUST and Dynamic FAUST [1, 2] (Figure 4), and realistic body models with clothing and accessories from 3DPW dataset [3] (Figure 5).

Figure 6 shows several examples of poses (Pose) and identities (Style) that the network sees while in training to illustrate the diversity of these two combinations. Figure 7 illustrates additional qualitative results on the FAUST [1] dataset of realistic characters. This example shows the consistency of shape transfer faced with several distinct poses and shapes. Equally, Figure 8 illustrate a challenging shape transfer of realistic clothed people (3DPW) [3] shape transfers. Note that the network at training only sees synthetic naked shapes (Figure 2), hence results on clothed people demonstrate the network generalization abilities.

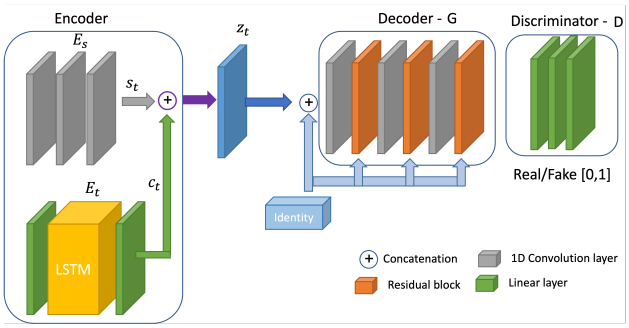

Figure 1: Network architecture: The spatial encoder E_s, considers input mesh vertex locations and yields a feature s_t per input frame. The temporal encoder E_t takes as input the vertex locations over all frames and outputs a hidden state vector c_t per frame. The decoder G produces the template mesh vertex locations for the new identity frames. The discriminator D predicts real or fake mesh shapes.

2. Limitations

While yielding state-of-the-art results on the motion transfer problem, our approach presents limitations with, for instance, input and output meshes with the same imposed topology. Another limitation lies in the improved robustness compared to a static learning-based strategy, which is yet not as good as with model-based strategies that can sometimes better preserve local shape structures, such as fingers.

Although not visible or encountered in the experiments, there is the possibility of foot skating issues as the current framework does not explicitly handle such a situation. Further, the collisions between body parts are not correctly expressed in the dataset, which induces the network to learn the wrong surface contacts, hence being a limitation of the current dataset. These are interesting problems to be investigated once the dataset correctly expresses body and feet to floor contacts.

*Institute of Engineering Univ. Grenoble Alpes

Table 1: Temporal Shape Transfer network architecture.

Encoder E_s Layers	Input Size	Output Size	Activation
1x1 Convolution	$3 \times x$	$64 \times x$	Leaky ReLU
1x1 Convolution	$64 \times x$	$128 \times x$	Leaky ReLU
1x1 Convolution	$128 \times x$	$1024 \times x$	Leaky ReLU

Encoder E_t Layers	Input Size	Output Size	Recur layers	Activation
Linear	$3 \times x$	1024	-	ReLU
LSTM	1024	256	3	-
Linear	256	$3 \times x$	-	tanh

Decoder G Layers	Input Size	Output Size	Activation
1x1 Convolution	$1030 \times x$	$1030 \times x$	Instance Norm
Adaptive ResBlock	$1030 \times x$	$1030 \times x$	-
1x1 Convolution	$1030 \times x$	$515 \times x$	Instance Norm
Adaptive ResBlock	$515 \times x$	$515 \times x$	-
1x1 Convolution	$515 \times x$	$257 \times x$	Instance Norm
Adaptive ResBlock	$257 \times x$	$257 \times x$	-
1x1 Convolution	$257 \times x$	$3 \times x$	Tanh

Discriminator D Layers	Input Size	Output Size	Activation
Linear	$3 \times x$	512	ReLU
Linear	512	128	ReLU
Linear	128	1	Sigmoid

References

[1] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014.

[2] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

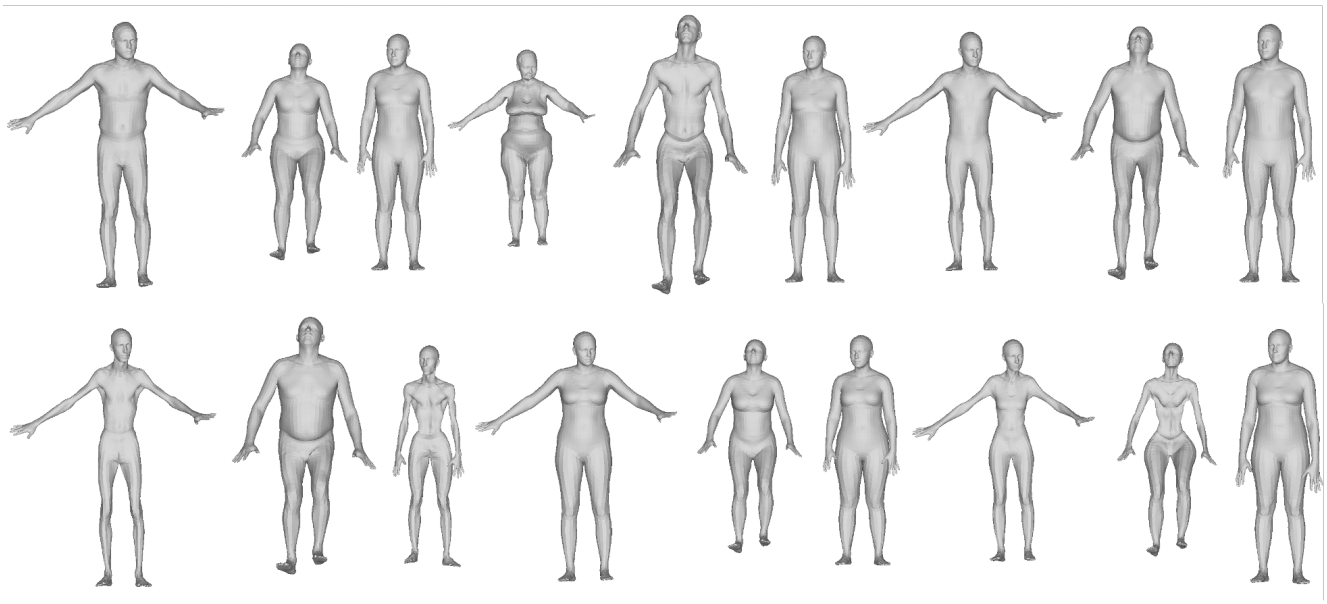

Figure 2: Shape identities used to train the proposed network. The syntetic shapes are a collection of female and male body types with realistic and unrealistic shapes.

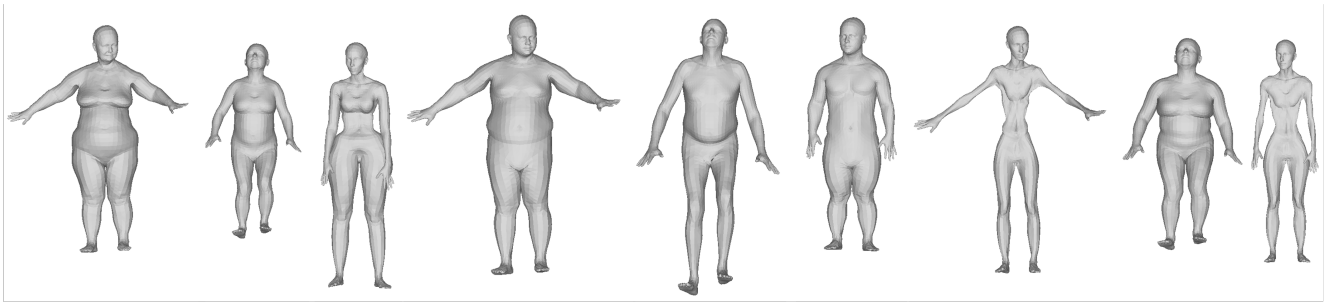

Figure 3: Shapes used as testing dataset. These examples were randomly generated and have examples which are far away from the training dataset, such as shapes with short or long legs as well as different variations of male and females body shapes.

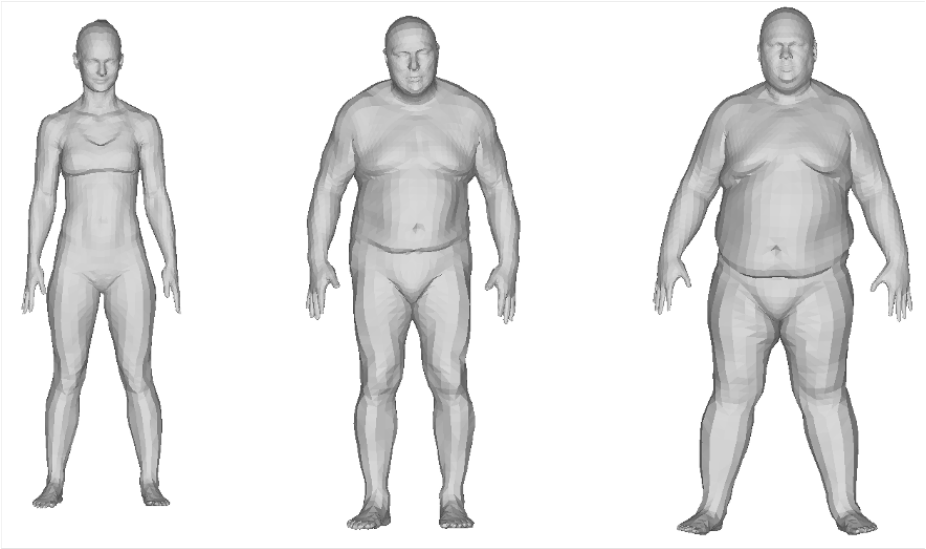

Figure 4: Realistic human bodies (FAUST and Dynamic FAUST [1, 2]) used to validate the proposed network and demonstrate generalisation ability to realistic individuals.

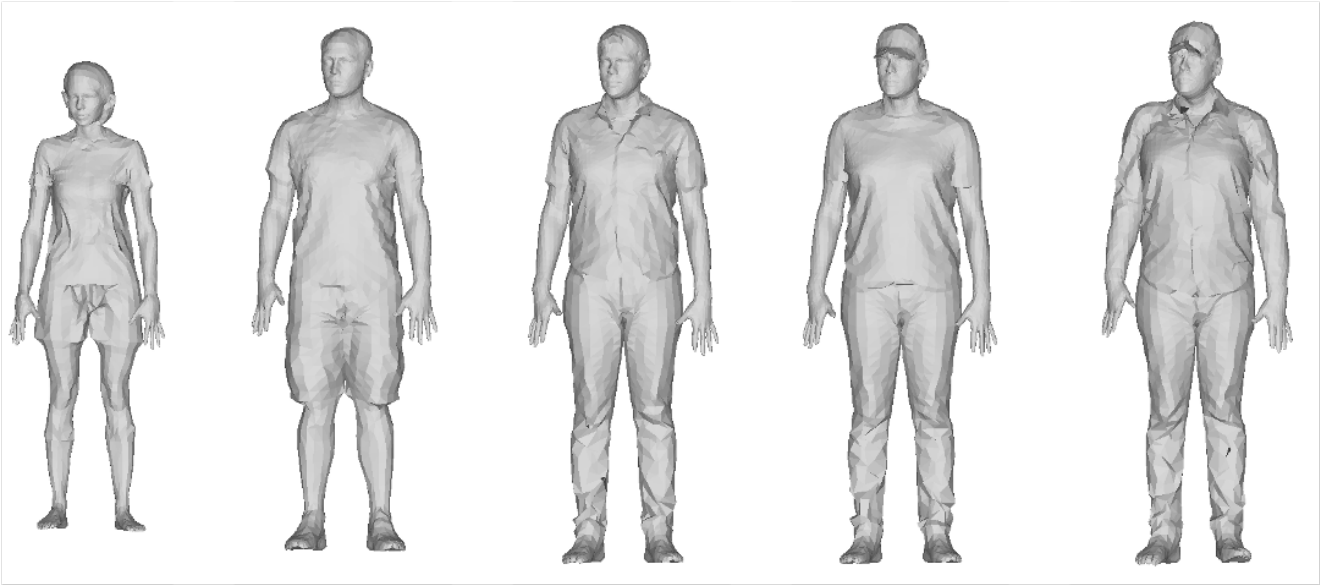

Figure 5: Realistic human bodies (3DPW [3]) used to validate the proposed network and demonstrate generalisation ability to clothed individuals.

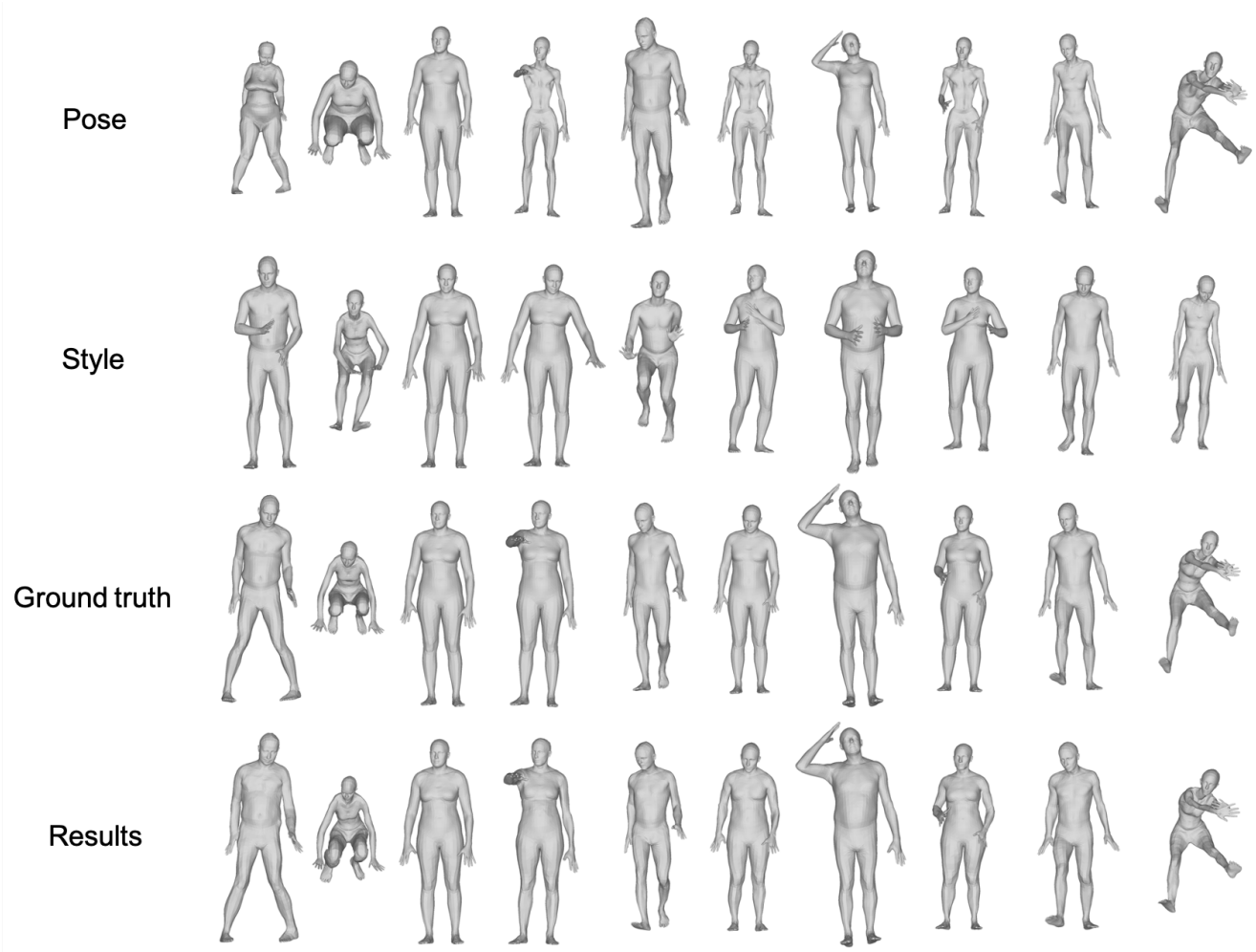

Figure 6: Synthetic models used to train the proposed network. Demonstrates the variety of pose and identity combinations while training, which is crucial to generalize to more challenging examples.

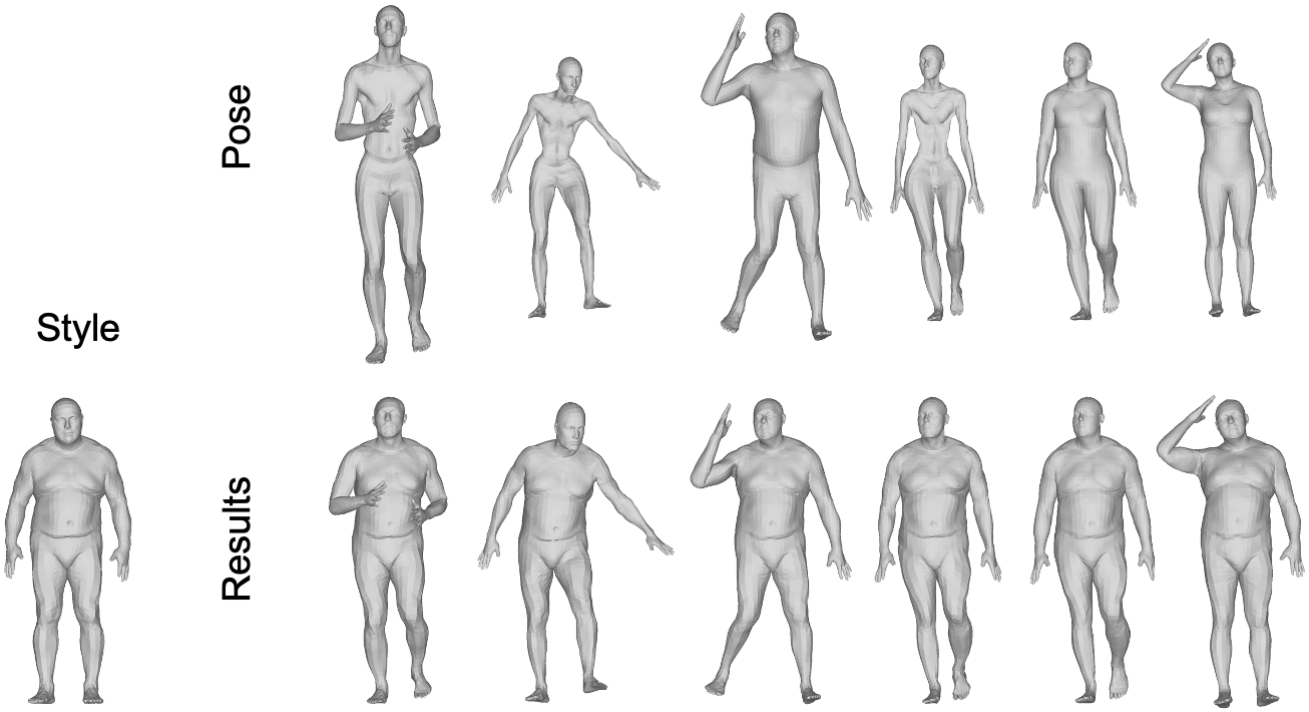

Figure 7: Realistic human body (FAUST [2]) transferred to different poses with varied body shapes.

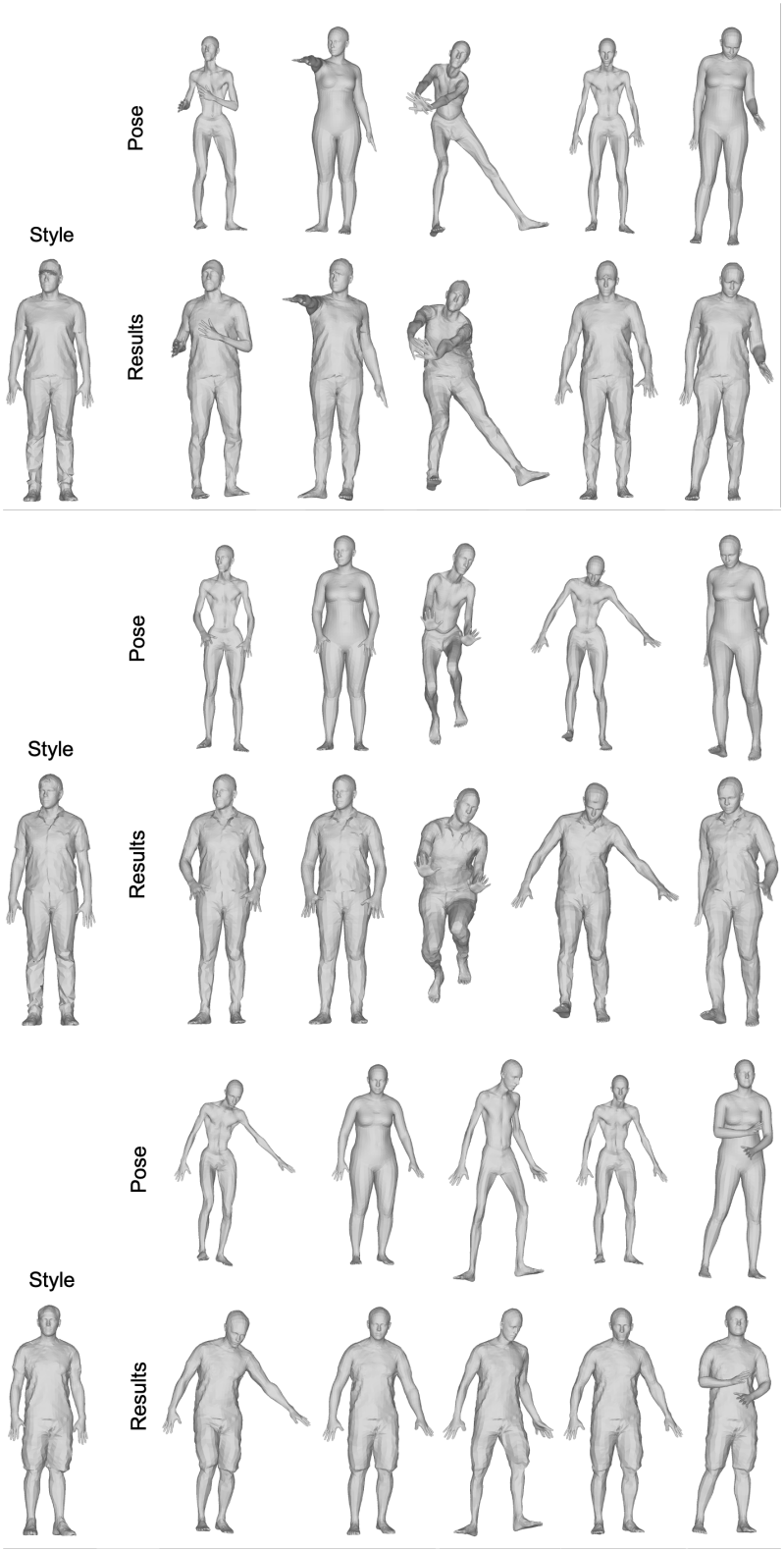

Figure 8: Realistic human bodies (3DPW [3]) and clothing transferred to different poses with varied body shapes.