Supplementary Material for Correspondence-Free Online Human Motion Retargeting

Rim Rekik[*1] Mathieu Marsot[*1] Anne-Hélène Olivier[2] Jean-Sébastien Franco[1] Stefanie Wuhrer[1]

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP [†], LJK, 38000 Grenoble, France

[2] Univ. Rennes, Inria, CNRS, IRISA, M2S, 35000 Rennes, France

[*] Joint first authorship: rim.rekik-dit-nekhili@inria.fr, mathieu.marsot@gmail.com

firstname.lastname@inria.fr

This supplementary material contains the network architecture and implementation details of our proposed framework and more details about the ablations. We further show comparisons of our method to a state-of-art retargeting method that uses correspondences, and provide additional explanations of the user study we conducted.

1. Network architecture and implementation details

Our method aims to preserve the motion of the source character at the skeletal level via a source motion retargeting model, and the detailed geometry of the target shape using a deformation model. The following provides details for each of the two parts.

In the following, N is the batch size, V is the number of vertices of scan $\mathcal{S}$, and J is the number of joints of skeleton $\mathcal{J}$.

All parts are implemented in PyTorch and optimized using Adam [1]. The training of the whole framework takes approximately one day on a GeForce RTX 2080 TI. The whole model has 14.6 million trainable parameters. The source motion retargeting model has 5 793 927 parameters, the 3D skeleton extraction model has 9 585 730 parameters and the deformation model has 257 564 parameters.

1.1. Source motion preservation

The source motion retargeting model is a reimplementation of [3]. We summarize the architecture in Table 1. FK denotes a forward kinematic layer.

The network is trained for 450 epochs where each epoch sees 2 316 examples. The learning rate start at 1e-4 and is divided by 10 every 150 epochs.

3D skeleton extraction Table 2 provides details on the architecture of the 3D skeleton extraction module.

[†]Institute of Engineering Univ. Grenoble Alpes

Index	Inputs	Operation	Output shape
(1)	Input	Joint positions $\mathcal{J}$	$N \times 3 \times J$
(1')	Input	Target skeleton $\mathcal{J}^B_{tpose}$	$N \times 3 \times J$
(2)	Input	Translations	$N \times 3$
(3)	(1) + (2)	GRU $(3 \times (J+1) \rightarrow 512, 2)$ Dropout $(p = 0.2)$	$N \times 512 \times J$
(4)	(3) + (1')	GRU $(3 \times J + 512 \rightarrow 512, 2)$ Dropout $(p = 0.2)$	$N \times 512$
(5)	(4)	Linear $(512 \rightarrow 6 \times J)$	$N \times 6 \times J$
(6)	(4)	Linear $(512 \rightarrow 3)$	$N \times 3$
(7)	(5) + (1')	FK	$N \times 3 \times J$

Table 1. The network architecture for source motion retargeting.

Index	Inputs	Operation	Output shape	Activation
(1)	Input	Source scan $\mathcal{S}^A$	$N \times 3 \times V$	-
(1')	(1)	backbone layer [4]	$N \times 4 \times 512$	-
(2)	(1')	Mean	$N \times 512$	-
(3)	(2)	Linear $(512 \rightarrow 256)$	$N \times 256$	Relu
(4)	(3)	Linear $(256 \rightarrow 64)$	$N \times 64$	Relu
(5)	(4)	Linear $(64 \rightarrow J \times 3)$	$N \times 3 \times J$	-

Table 2. The network architecture for 3D skeleton extraction.

The network is trained for 34 epochs where each epoch sees 65 000 examples. The learning rate is 1e-4 for the first 30 epochs then 1e-5 for 2 epochs and 1e-6 for 2 epochs.

1.2. Target geometry preservation

The architecture of the deformation module is composed of the skinning weights predictor network detailed in Table 3 and the pose corrective offset predictor network detailed in Table 4.

Index	Inputs	Operation	Output shape	Activation
(1)	$N \times 3 \times (J+1)$	Linear $(3 \times (J+1)) \rightarrow 256)$ Dropout $(p = 0.2)$	$N \times 256$	Relu
(2)	(1)	Linear (256→256) Dropout $(p = 0.2)$	$N \times 256$	Relu
(3)	(2)	Linear $(256 \rightarrow J)$	$N \times J$	Softmax

Table 3. The network architecture for skinning weights predictor.

Index	Inputs	Operation	Output shape	Activation
(1)	$N \times 3 \times (J+1)$	Linear $(3 \times (J+1) \rightarrow 256)$ Dropout $(p = 0.2)$	$N \times 256$	Relu
(2)	(1)	Linear (256→256) Dropout $(p = 0.2)$	$N \times 256$	Relu
(3)	(2)	Linear $(256 \rightarrow 3)$	$N \times 3$	-

Table 4. The network architecture for the pose corrective offset predictor.

The skinning weights predictor and pose corrective offset predictor networks are trained jointly for 28 epochs where each epoch sees 65 000 examples. For the first epoch, only the skinning weights predictor network is optimized. The learning rate is 1e-3.

2. Comparison to retargeting method using correspondences, TST

Table 5 quantitatively compares our method to TST [2], a state-of-the-art deformation transfer method that requires point-to-point correspondence. Note that while TST performs better than our method, the improvement is minor. This is a strong result showing that our correspondence-free method only gets slightly outperformed by a state-of-the-art method that relies on additional correspondence information.

	Skeletal motion				Detail preserv.		
	MPJPE (m) ↓	PA-MPJPE (m) ↓	Acc ↓	PA-Acc ↓	MPVD (m) ↓	PA-MPVD (m) ↓	MDEL (mm) ↓
Naked target shapes from SMPL test set							
TST [2]	**0.152**	**0.028**	**0.005**	**0.004**	**0.130**	**0.028**	0.866
Ours	0.158	0.043	0.021	0.008	0.134	0.040	**0.524**
Clothed target shapes from CAPE test set							
TST [2]	**0.096**	**0.027**	**0.005**	**0.004**	**0.064**	**0.028**	0.953
Ours	0.105	0.040	0.023	0.008	0.075	0.038	**0.539**

Table 5. Comparison of our correspondence-free method to a state-of-the-art method which needs correspondences. The comparison is done on naked (top) and clothed (bottom) target shapes.

3. Further ablation studies for source motion preservation

We now present further ablation studies for the source motion retargeting model.

First, Table 4 in the main paper shows that using $\mathcal{L}_{rot}$ and the 6D representation lead to significant improvements on all quantitative metrics. Figure 1 further shows an example retargeting with different rotation representations and with and without the cycle consistency loss. Note that our changes significantly improve the retargeting result.

Second, we show quantitatively in Table 6 that adding $\mathcal{L}_{smooth}$ improves the source motion retargeting. We evaluate the skelatal motion since the $\mathcal{L}_{smooth}$ operates on skeletons.

	Skeletal motion			
	MPJPE (m) ↓	PA-MPJPE (m) ↓	Acc ↓	PA-Acc ↓
L_{motion} without $\mathcal{L}_{smooth}$	0.160	0.044	0.023	0.008
L_{motion} with $\mathcal{L}_{smooth}$	**0.158**	**0.043**	**0.021**	**0.008**

Table 6. Quantitative ablation of $\mathcal{L}_{smooth}$ on SMPL shapes.

3D skeleton extraction In this paragraph, we study the effect of different point cloud encodings on the 3D skeleton extraction model. The main paper presents quantitative results in favour of using PointFormer over PointNet for this purpose. Figure 2 shows a comparison on a challenging example. When using PointNet, the torso prediction is far from the ground truth while using PointFormer, the only noticeable errors are on the right foot and wrist joints.

4. User study

In order to further assess our method, we designed a user study to evaluate the perceived resulting motion with respect to the source motion, and the perceived resulting shape with respect to the target shape.

4.1. Participants

30 unpaid participants volunteered for the experiment (14 females, 16 males; age: average=32±11, range 21-64). They had different expertise levels in animation (median = 2, interquartile range (IQR) = 3, range 1-7 using a Likert-scale from 1 (novice) to 7 (expert)) and human motion (median = 3, IQR = 4, range 1-7 using a Likert-scale from 1 (novice) to 7 (expert)). They all had normal or corrected to-normal vision, and gave written and informed consent. The study conformed to the declaration of Helsinki, and was approved by our institutional ethics committee.

4.2. Design

For each trial, participants were seated in front of a 24" computer screen and were presented stimuli showing a source motion, a target shape, and a retargeting result, and asked to rate motion and shape similarity, respectively, with source motion and target shape on a 7-point Likert scale, as illustrated on Figure 3.

We built the experiment using PsychoPy software. We used a within-subject design with retargeting methods as a main factor. We considered 3 retargeting methods: ours and

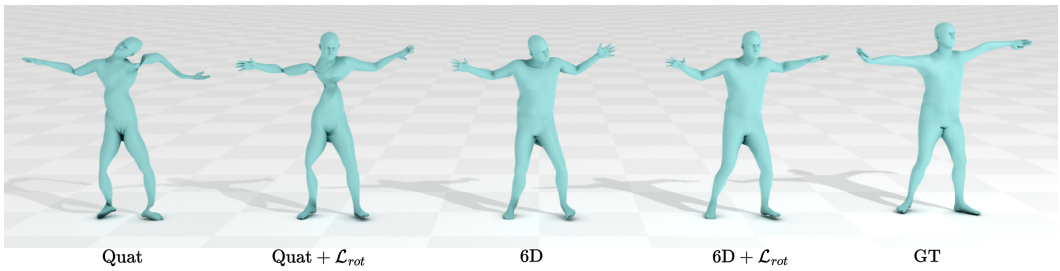

Figure 1. Ablation of SMRM. Retargeting result on a challenging HipHop motion from a female to a male body shape. Quaternions are prone to twist, introducing $\mathcal{L}_{rot}$ improves the head and feet retargeting.

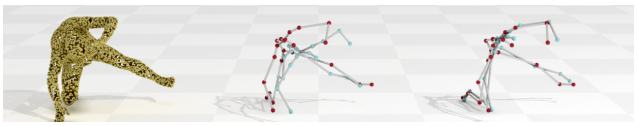

Figure 2. Ablation of 3D Skeleton extraction: with PointNet (middle) and PointFormer (right) on a challenging pose from the AMASS test set (right). Ground truth shown in red, regressed results in blue.

H4D (without correspondences), and TST (with correspondences). 15 repetitions were performed for each method using different motions and shape transfers. We selected 3 types of motions, shown in Figure 4:

1. walking, which involves a global displacement with cyclic body-motions,
2. throwing, which is a more local and discrete upper-body motion, and
3. hip-hop, which is a complex motion.

Figure 5 shows the selected shape transfers including gender, body anthropometrics and clothes changes, namely:

1. female to male,
2. male to female,
3. corpulent to skinny,
4. skinny to corpulent, and
5. naked to dressed.

After 4 training trials, participants performed in total 45 trials. Trials were presented by block of shape transfer. At the beginning of each block, participants were shown the T-pose of the source body shape and the target-body shape. We used a latin-square design to randomize block orders. Then, for each block, 9 trials (3 methods $\times$ 3 motions) were performed by participants in a randomized order. For each participant and each trial, dependent variables were the score of similarity regarding motion and shape. Statistical analysis was performed in R using repeated-measures ART-ANOVAs and post-hoc Tukey tests. The level of significance was set to 0.05.

4.3. Results

Our results show a main effect of the retargeting model both for the perceived motion ($F(2, 58) = 297.9, p < 0.0001$) and the shape similarity ($F(2, 58) = 85.3, p < 0.0001$). The effect is important since the effect size η_p^2 was respectively 0.91 and 0.74. As illustrated in Figure 6, post-hoc tests show that our method was better rated than H4D both for motion ($t(58) = 19.01, p < 0.0001$) and shape similarity ($t(58) = 12.74, p < 0.0001$) in comparison to the source motion and target shape. In addition, while our method had lower scores than TST for the perceived motion ($t(58) = 3.75, p < 0.001$), it outperformed TST for the perceived shape ($t(58) = 8.87, p < 0.0001$).

5. Potential negative societal impact

We present a method for long-term and geometrically detailed motion retargeting between different digitized human models. It could be used without the consent of the user to animate static 3D scans, or even 3D reconstructions generated from 2D images, *e.g.* to generate disinformation.

References

[1] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 1

[2] João Regateiro and Edmond Boyer. Temporal shape transfer network for 3d human motion. In *International Conference on 3D Vision*, 2022. 2

[3] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *Conference on Computer Vision and Pattern Recognition*, pages 8639–8648, 2018. 1

[4] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 1

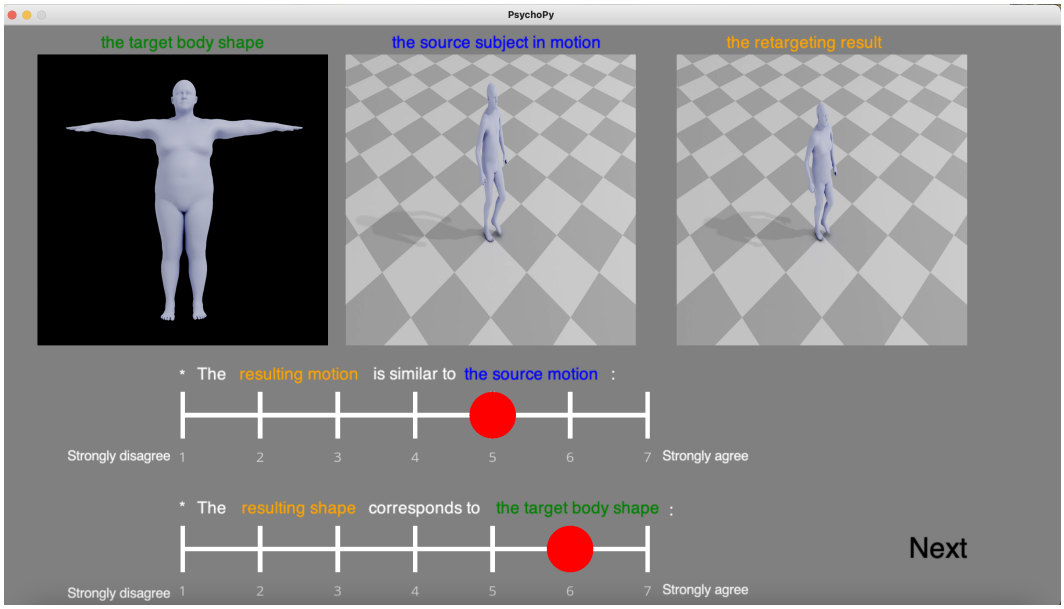

Figure 3. Screenshot of the stimuli and questions presented to participants during the user study.

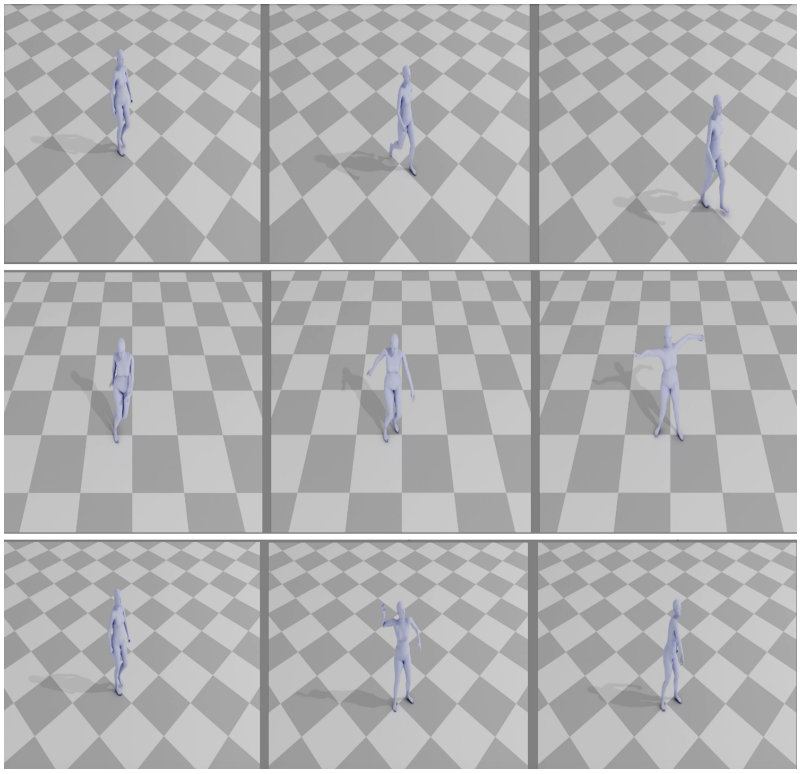

Figure 4. Illustration of the 3 motions selected for the user study. Top: Walking motion, middle: Hip-Hop motion and bottom: Throwing motion. Those examples were used as source motion for the female shape.

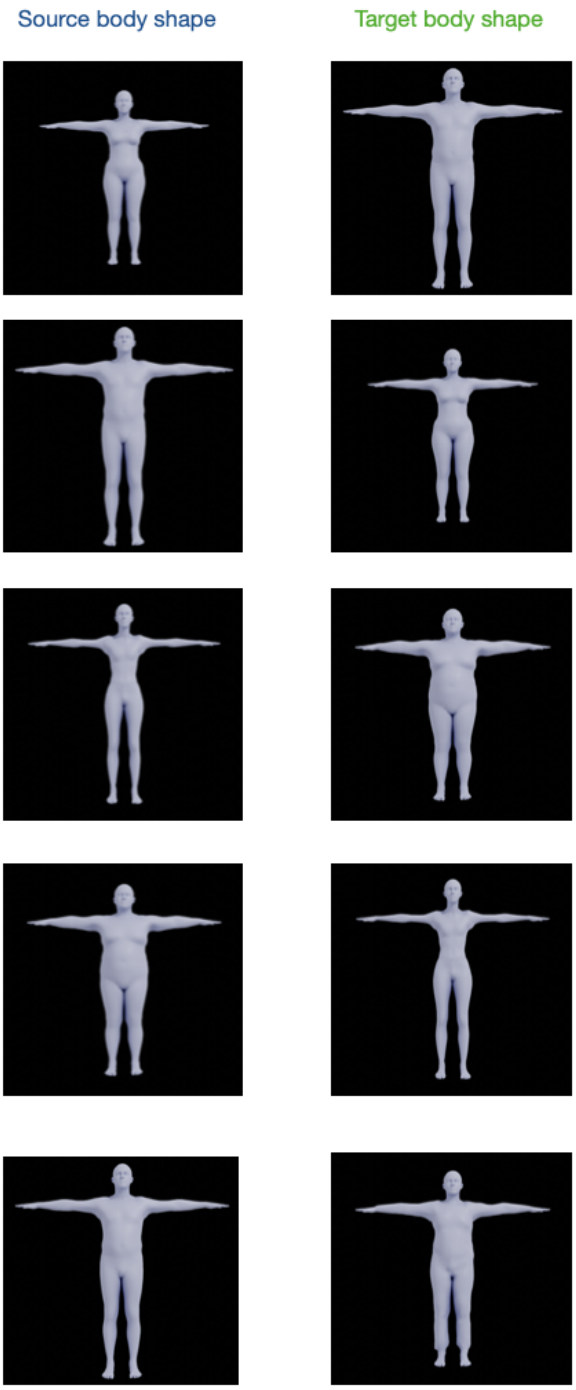

Figure 5. Illustration of the 5 shape transfers selected for the user study. Source shape is on the left, target shape is on the right.

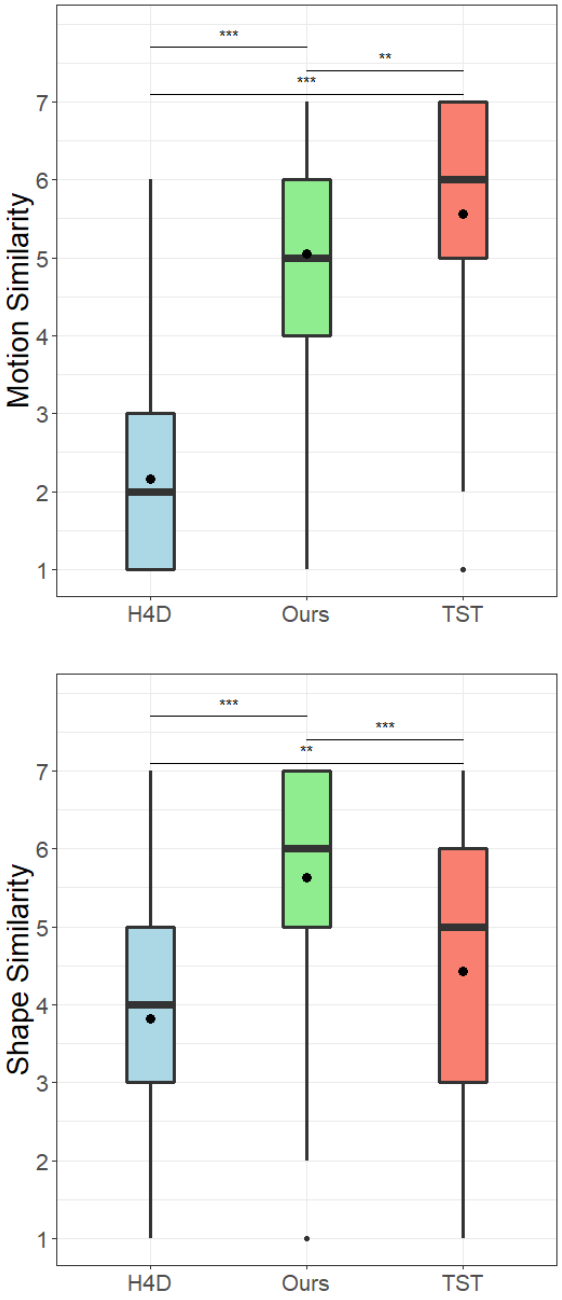

Figure 6. Results of the user study. Top: boxplots illustrating scores given by participants for motion similarity between source motion and retargeting result. Bottom: boxplots illustrating scores given by participants on shape similarity between target shape and retargeting result. Stars symbols highlight significant differences between models from the post-hoc tests ($** : p < 0.001, *** : p < 0.0001$).