Correspondence-free online human motion retargeting

Rim Rekik, Mathieu Marsot, Anne-Hélène Olivier, Jean-Sébastien Franco, Stefanie Wuhrer

Correspondence-free online human motion retargeting

Mathieu Marsot[*1] Rim Rekik[*1] Stefanie Wuhrer[1] Jean-Sébastien Franco[1]
Anne-Hélène Olivier[2]
[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: mathieu.marsot@inria.fr, rim.rekik-dit-nekhili@inria.fr

Abstract

We present a novel data-driven framework for unsupervised human motion retargeting which animates a target body shape with a source motion. This allows to retarget motions between different characters by animating a target subject with a motion of a source subject. Our method is correspondence-free, i.e. *neither spatial correspondences between the source and target shapes nor temporal correspondences between different frames of the source motion are required. Our proposed method directly animates a target shape with arbitrary sequences of humans in motion, possibly captured using 4D acquisition platforms or consumer devices. Our framework takes into account long-term temporal context of* 1 *second during retargeting while accounting for surface details. To achieve this, we take inspiration from two lines of existing work: skeletal motion retargeting, which leverages long-term temporal context at the cost of surface detail, and surface-based retargeting, which preserves surface details without considering long-term temporal context. We unify the advantages of these works by combining a learnt skinning field with a skeletal retargeting approach. During inference, our method runs online,* i.e. *the input can be processed in a serial way, and retargeting is performed in a single forward pass per frame. Experiments show that including long-term temporal context during training improves the method's accuracy both in terms of the retargeted skeletal motion and the detail preservation. Furthermore, our method generalizes well on unobserved motions and body shapes. We demonstrate that the proposed framework achieves state-of-the-art results on two test datasets. Our code is available* https://gitlab.inria.fr/rrekikdi/human-motion-retargeting2023.

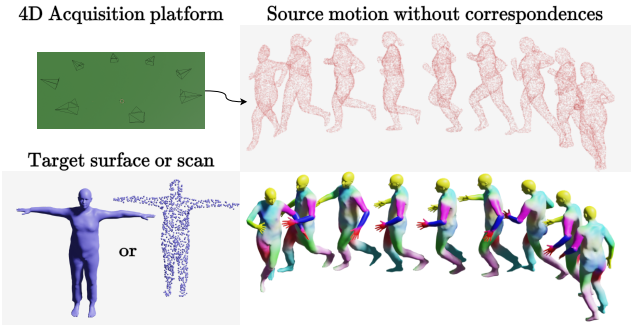

Figure 1. Given an untracked source motion (top) and a target body shape (bottom left), our method animates the target with the source motion, preserving temporal correspondences of the output motion (bottom right).

1. Introduction

Human motion retargeting is the process of animating a target character with the motion sequence of a source character. We study this problem for densely sampled 3D surfaces, and arbitrary sequence duration.

Applications include video gaming, movie making, avatar animation, and augmenting existing datasets of 4D body motions *e.g.* [5, 6, 27, 38]. The goal of this work is correspondence-free retargeting that takes as input unstructured 4D data for which neither correspondences between the source and target shape nor correspondences between different frames of the source motion are given. This allows to directly animate a human target shape with arbitrary sequences of humans in motion, possibly captured using 4D acquisition platforms. Fig. 1 illustrates the problem's input and output.

Motion retargeting has recently made significant progress and leads to impressive results. Skeleton-based methods allow for retargeting while taking long-term temporal context of about 2 seconds into account *e.g.* [14, 22, 37]. Surface deformation-based methods allow to retarget

geometric details while possibly capturing short time dynamics of motion *e.g.* [4, 33, 34, 39].

Most of these retargeting works take as input structured data in the form of skeletons or template-aligned surfaces for which correspondence information is known. Solving the retargeting problem for unstructured 3D data is challenging as computing correspondences is a combinatorial problem.

Recently, a first solution for correspondence-free motion retargeting has been proposed [17]. This work introduces a generic motion prior to describe how body shapes move in a low dimensional space, and takes long-term temporal context into account. While this opens the door to novel solutions, during inference the method is limited to retarget sequences of fixed length. Inspired by this work, our framework also learns long-term temporal context.

In this work, we propose the first correspondence-free method for online motion retargeting that learns long-term temporal context. Especially, we are able to learn context from $1s$ of motion sequences, which improves the accuracy of motion retargeting even with challenging examples. We achieve this by combining the advantages of skeletal motion retargeting and surface-based motion retargeting to learn about both temporal context and geometric detail. More specifically, we combine skeletal motion retargeting and automatic skinning, thereby allowing to handle high-dimensional spatio-temporal data. The reason is that skeletal retargeting methods achieve good results by encoding long-term temporal context [37] and that recently, methods on automatic skinning have been proposed that we can include in our inference model [24, 30, 40, 41]. Our hypothesis is that the locomotion information we want to extract from the source sequence is explained at a skeletal level, while the dense surface details are intrinsic to the target shape.

In practice, our method is divided in three modules. First, a skeleton regressor allows to transition between unstructured 3D point clouds and skeletal representations. Second, a skeleton-based motion retargeting method allows to transfer the source motion to the target skeleton. Third, an automatic skinning predictor, which combined with classical linear blend skinning (LBS) reposes the unstructured target point cloud.

Our main contributions are summarized below:

- We propose the first correspondence-free online approach for dense human motion retargeting that learns temporal context.
- We demonstrate that long-term temporal context improves the accuracy of motion retargeting.
- We demonstrate state-of-the-art results for both geometric detail preservation and skeletal-level motion retargeting.

2. Related Work

Existing works addressing the motion retargeting problem can be categorized into three lines of work: works that operate on the skeleton level, works that transfer shape deformations at a dense geometric level, and motion priors, which can be applied to motion retargeting.

2.1. Skeletal motion transfer

Early works for motion retargeting are based on skeleton parametrizations, where human pose is described with a sparse set of joint locations. Gleicher *et al.* [14] introduced this task, considering it as an optimization problem with kinematic constraints over the entire motion sequence. The resulting skeletal parametrizations can then be animated with manually computed skinning weights [23, 42] and using blending techniques [18, 20].

Follow up works [11, 19, 21, 35] also require iterative optimization with hand-designed kinematic constraints for particular motions. With the surge accessibility of captured motion data and the efficiency of deep learning techniques, new data-driven approaches [1, 13, 15, 37] showed outstanding results without requiring many handcrafted energies. Some of these data-driven approaches [13, 15] require paired training data, whose design involves human effort. Therefore, another line of works [1, 37] propose unsupervised retargeting strategies. Villegas *et al.* [37] propose an unsupervised motion retargeting framework based on adversarial cycle consistency to ensure plausibility of the retargeted motions. This method generates natural motions for unseen characters, but only operates at a skeletal level and does not include geometric details.

More recent work [36] proposes to include geometry and investigates hybrid skeleton-based motion retargeting with both a data-driven network and a post inference optimization to preserve self-contacts and prevent interpenetration. This method shows outstanding results, yet requires manually handcrafted skinning weights as additional input.

2.2. Shape deformation transfer

Shape deformation transfer methods operate directly on the surface either by using representations of 3D body shape that disentangle shape and pose [12, 44] or by optimising mesh deformations [3, 7, 34, 45]. Some works [3, 34] consider motion retargeting as pose deformation transfer, encoding the pose of the source character and transferring the deformation of the associated mesh to the target. Other works [4] consider motion retargeting as shape deformation transfer, encoding the shape identity of the target character and transferring it to the source character. These works showed impressive results but always operate on meshes in correspondence. As providing spatio-temporal correspondences between source and target meshes is a difficult task,

Wang *et al*. [39] propose a new style transfer approach that learns to transfer a character's shape style onto another posed character while learning correspondences between point clouds with different structures. This work shows good results for per-frame pose transfer but suffers from stretching artefacts in case of body contacts or limbs in proximity. To remove these artefacts, recent works add temporal context by considering $3-4$ consecutive frames. Chen *et al*. [8] include temporal information using a sequence-to-sequence architecture that operates on unstructured 4D data, by requiring an entire target sequence instead of a single target frame. Regateiro *et al*. [33] use a recurrent neural network that extracts the temporal information from the source motion, which leads outstanding results.

These deformation-based approaches cannot be extended to learn long-term temporal context beyond a few consecutive frames due to computational complexity. We show experimentally that our method performs almost on par with Regateiro *et al*. [33], without having access to correspondence information.

2.3. Generic motion priors

Generic motion priors have recently been applied successfully to motion retargeting [17, 28]. These priors represent dense human motion using a latent representation that disentangles the locomotion information and the dense body shape information. They allow to retarget the body motion to different body shape while considering long-term temporal context. Marsot *et al*. [28] operate with structured data and show that the prior learns correlations between dense body shape and skeletal locomotion. Closer to our setting, Jiang *et al*. [17] consider unstructured point clouds and show that their method is able to retain dense surface details of the target shape.

While motion priors can account for long-term temporal context, they cannot easily be applied to sequences of arbitrary duration at inference because they operate on normalized temporal segments of motion. We demonstrate experimentally that our method outperforms Jiang *et al*. [17].

2.4. Positioning

Table 1 summarizes the positioning of our work w.r.t. state-of-the-art. We classify approaches based on four criteria: using long-term temporal context ($0.5s$ or more) for training, allowing for online inference, modeling geometric detail, and operating on unstructured data for which no correspondences are known. By combining the advantages of skeleton-based retargeting and skinning fields, we propose the first correspondence-free retargeting approach that learns long-term temporal context, allows for arbitrary duration at inference, all while modeling geometric detail at the surface level.

Method	Temporal context	Online inference	Geometric detail	Unstructured
Skeleton-based				
[1, 2, 22, 37]	✓	✓	✗	✗
[36]	✓	✓	✓	✗
Shape deformation transfer				
[4, 9, 33]	✗	✓	✓	✗
[8, 39]	✗	✓	✓	✓
Motion priors				
[28]	✓	✗	✓	✗
[17]	✓	✗	✓	✓
Ours	✓	✓	✓	✓

Table 1. Positioning w.r.t. state-of-the-art retargeting approaches. We propose the first correspondence-free retargeting approach that learns long-term temporal context, allows for arbitrary duration at inference, all while modeling geometric detail at the surface level.

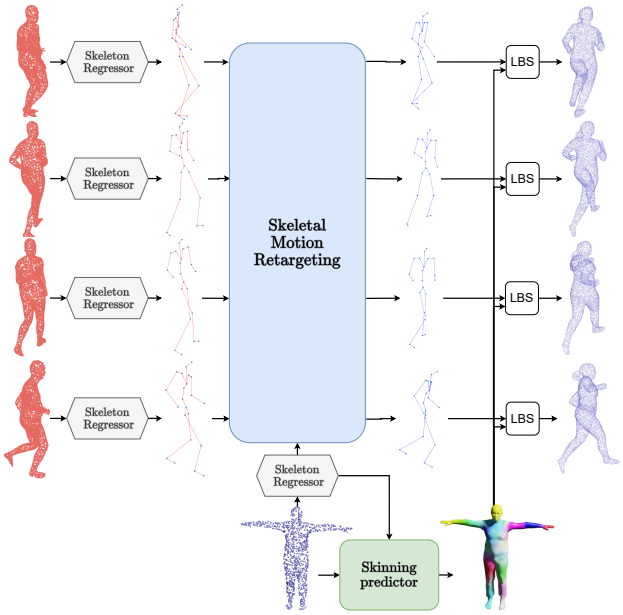

Figure 2. Our method takes a source sequence of unstructured point clouds along with a target point cloud as input and outputs the target character performing the input motion. The method proceeds in three stages: the first one (grey boxes) extracts per-frame skeletal representations from the input source sequence, the second one (blue box) retargets the locomotion to the target character at the skeleton level, and the third one (green box) adds the surface details of the target character to the resulting motion using a predicted skinning field.

3. Motion retargeting model

In this section, we introduce notations, give an overview of our retargeting model, and provide details for all its stages. We then discuss our implementation choices and

training strategy.

3.1. Overview

The input source motion is characterized by a sequence of n point clouds $\{\mathcal{S}_i^A\}_{i=1}^n$ without temporal correspondences and the target shape B by a single point cloud, possibly with connectivity information, of B in T-Pose $\mathcal{S}_{tpose}^B$. Our objective is to generate the sequence of retargeted scans $\{\mathcal{S}_i^B\}_{i=1}^n$, where $\mathcal{S}_i^B$ imitates the pose of $\mathcal{S}_i^A$ and is in correspondence with $\mathcal{S}_{tpose}^B$. Fig. 2 gives a visual overview of our method.

A common strategy for motion retargeting is to disentangle the high-level motion from the shape deformation caused by the body shape of the source $\mathcal{S}^A$. To facilitate this task, we make the hypothesis that the detailed geometry of the source does not contain information about its high-level motion, which is fully explained at a skeletal level. This justifies our choice for a three-step framework for motion retargeting.

First, we extract a sequence of skeletons from $\{\mathcal{S}_i^A\}_{i=1}^n$. To obtain a skeletal representation from correspondence-less point clouds, we regress a set of joint positions with a Skeleton Regressor module (SKR). This skeletal parameterization has three major advantages. It provides correspondence information between input frames, drastically reduces the dimensionality of the data and fully encodes the shape induced variability in bone length.

Second, from the resulting skeletal parameterization of the source $\{\mathcal{J}^A\}_{i=1}^n$, we extract its high level motion information and retarget this motion to the skeletal parameterization of the target $\mathcal{J}_{tpose}^B$. To do so, we use a context-aware skeletal motion retargeting model (SMRM) [37] which given $\{\mathcal{J}^A\}_{i=1}^n$ and $\mathcal{J}_{tpose}^B$ outputs the relative joint rotations $\{q^B\}_{i=1}^n = \text{SMRM}(\{\mathcal{J}^A\}_{i=1}^n, \mathcal{J}_{tpose}^B)$ that can be applied to $\mathcal{J}_{tpose}^B$ using forward kinematics to obtain the retargeted skeletal parameterization of B called $\{\mathcal{J}_i^B\}_{i=1}^n$.

Third, we recover shape details of B to compute the final animation. As B is a point cloud, we learn a dense skinning prior SKIN, which given a skeleton $\mathcal{J}$, associates to every 3D point of B one skinning weight per joint $W = \text{SKIN}(x, y, z, \mathcal{J})$. Given the skeletal representation $\mathcal{J}_{tpose}^B$, the set of rotations $\{q^B\}_{i=1}^n$ and the skinning weights W^B of all point of B, we generate the retargeted sequence using linear blend skinning (LBS) as $\{\mathcal{S}_i^B\}_{i=1}^n = \text{LBS}(\mathcal{J}^B, \{q^B\}_{i=1}^n, W^B)$.

Our three-stage framework allows to solve the motion retargeting problem for dense geometric data, where each frame consists of thousands of vertices, while taking into account long-term temporal context. A key advantage of our framework over temporal shape deformation transfer methods [8, 33] is its low computational complexity for increasing sequence duration considered during training. The reason is that point clouds of the source and target shapes are processed by considering a single frames at a time in the first and third stages, while the long-term motion is considered at the skeletal level during the second stage. The following provides details about each of these three parts.

3.2. Skeleton Regressor

Our first objective is to extract a skeleton from a point cloud. We parameterize the skeleton using 22 joint positions, which corresponds to the joints of the SMPL body model [25], a commonly used parametric model of naked human bodies.

To operate on unstructured point clouds, the model should be robust w.r.t. the number and order of the observed points. To achieve this task, we use a PointNet [32] based neural regressor to extract order-invariant features from the point cloud, followed by a multi layer perceptron to regress the joint positions from these features.

3.3. Skeletal motion retargeting

The second module of our method retargets the source skeletal motion to the target skeleton. This module needs to properly extract long-term motion features. Practically, the locomotion is characterized by the evolution of the skeleton relative joint positions over time and the evolution of a global displacement vector characterizing the position of its root joint in the world coordinate frame. As ground truth pairings between a motion and its retargetted version are not available during training, we train in an unsupervised setting by leveraging prior works [37].

That is, we start by extracting high level locomotion properties from a source sequence of joint positions $\{\mathcal{J}_i^A\}_{i=1}^n$ using a first recurrent network. These features are then leveraged along with the joints of a target skeleton in T-pose $\mathcal{J}_{tpose}^B$ to generate joint rotations in quaternion representation $\{q_i^B\}_{i=1}^n$ using a second recurrent unit. Finally, these rotations are applied to the target skeleton using a differentiable forward kinematics layer to generate the retargeted skeletal motion $\{\mathcal{J}_i^B\}_{i=1}^n$.

3.4. Skinning predictor

The third stage takes as input character B in a canonical T-pose along with its T-pose skeletal representation, and animates the B using the joint rotation $\{q_i^B\}_{i=1}^n$ predicted by the skeletal motion retargeting module. A common stategy to animate 3D shapes given their skeletal motion is to associate surface vertices to the skeleton joints by a set of skinning weights and to animate the shapes using skinning techniques such as linear blend skinning. Recently, works in the area of neural human skinning [10, 30, 31, 40, 41] have succeeded in automatically predicting skinning weights given an input mesh. Our setting differs in two aspects. First, we dispose of a T-pose skeleton of the target shape, predicted by our skeleton regression module and second, our method

operates on point clouds with different structures instead of meshes with fixed topology.

Inspired by neural skinning methods, we learn a skinning predictor by characterizing a dense skinning prior over 3D space given a skeleton in T-pose. This predictor is a mapping function from $\mathbb{R}^3$ to $\mathbb{R}^{22}$. It associates each 3D point to a set of 22 skinning weights W using a multi layer perceptron. Learning a continuous skinning field in $\mathbb{R}^3$ allows to handle unstructured point clouds.

To ensure that the model is translation invariant, it takes as input an ordered vector of distances between a queried 3D point and each joint of the skeleton in T-pose. Inspired by [10], we also constrain the predicted weights to sum to 1 using a softmax activation.

3.5. Training

We choose a stage-wise training strategy in order to guarantee the training stability and reduce the computational complexity. Recent works used similar training strategies for related problems and demonstrated impressive results [16, 17, 29, 43].

To train all three parts, we use AMASS dataset [27], which contains a collection of motion capture datasets that have been fitted by the parametric body model SMPL to obtain dense per-frame representations [25]. As training data, we consider a subset of 120 body shapes, seen performing 2536 different motions for a total of 65000 frames. The motion sequences are all temporally aligned to 30FPS. The 3D shape is presented by 6890 unordred vertices and the skeletal parametrization is presented by 22 joint positions.

For the skeleton regressor and skinning predictor modules, we train in a static setting. For the skeletal motion retargeting module, we train on randomly sampled subsequences with a fixed duration.

Skeleton regressor For the skeleton regressor, we use for training static 3D meshes provided by AMASS with their corresponding SMPL skeletons. To avoid learning the bias of the SMPL template mesh, which provides correspondence information, we randomly uniformly sample N points on the surface of each 3D mesh to generate ground truth scans $\mathcal{S}_{GT}$. During training, we minimize as loss the mean squared error (MSE) between the ground truth SMPL joints $\mathcal{J}_{GT}$ and the predicted joints:

$$\mathcal{L}_{\text{SKR}} = MSE(E(\mathcal{S}_{GT}), \mathcal{J}_{GT}), \tag{1}$$

where $E(\mathcal{S}_{GT})$ is the skeleton extracted from point cloud $\mathcal{S}_{GT}$ by our network E. We experimented the addition of a bone length preservation loss but it did not lead to any noticeable improvement in the retargeted motions.

Skeletal motion retargeting To train this module, we randomly sample 30 frames sequences which we found to be the optimal context duration according experiments in Table 4. To train the model in an unsupervised setting, Villegas *et al*. [37] propose a cycle-consistency loss and an adversarial loss. A smoothing loss on the global acceleration is also minimized. The adversarial term leverages an additional discriminator network which is trained to differentiate between real and retargeted motions. Both the discriminator and the motion retargeting network are trained in a min-max game where the motion retargeting network tries to fool the discriminator. The smoothing loss minimizes the mean acceleration of the retargeted motion.

To further regularize the generated motions, we extend the joint based cycle-consistency loss of [37] with a rotation based cycle consistency to prevent unrealistic motions.

To compute the cycle consistency loss, the motion of character A, $\{\mathcal{J}_i^A\}_{i=1}^n$, is retargeted to character B, yielding $\{q_i^B, \mathcal{J}_i^B\}_{i=1}^n$. The retargeted motion is then retargeted back to A, yielding $\{\hat{q}_i^A, \hat{\mathcal{J}}_i^A\}_{i=1}^n$. The joint based cycle consistency loss evaluates the mean squared error between $\{\mathcal{J}_i^A\}_{i=1}^n$ and $\{\hat{\mathcal{J}}_i^A\}_{i=1}^n$ and our added cycle consistency loss on rotation evaluates the mean squared error between $\{q_i^A\}_{i=1}^n$ and $\{\hat{q}_i^A\}_{i=1}^n$.

The complete loss for the retargeting module is

$$\begin{aligned} \mathcal{L}_{\text{SMRM}} &= \mathcal{L}_{cycon} + \mathcal{L}_{adv} \\ &+ \lambda_{rot}\mathcal{L}_{rot} + \lambda_{smooth}\mathcal{L}_{smooth} \end{aligned} \tag{2}$$

where $\lambda_{rot}, = 0.01$ and $\lambda_{smooth} = 0.001$ are fixed weighs that balance the influence of the losses. We show experimentally the benefit of adding $\mathcal{L}_{rot}$ in Table 2.

Skinning predictor The skinning predictor animates a target scan $\mathcal{S}_{tpose}$ using a set of rotation q and a T-pose skeleton $\mathcal{J}_{tpose} = \text{SKR}(\mathcal{S}_{tpose})$ This module is trained using pairs of scans $\{\mathcal{S}_{tpose}, \mathcal{S}\}$ of a same person in correspondence and with rotations q that explain the pose of $\mathcal{S}$. To preserve correspondences while removing the bias due to SMPL topology, the scans are generated by uniformly sampling the ground truth meshes using precomputed barycentric coordinates from the vertices of SMPL.

The skinning predictor is modelled with a three layer MLP followed by a softmax activation. To train this network we minimize the loss :

$$\mathcal{L}_{\text{SKIN}} = \sum_{(p,p') \in (\mathcal{S}_{tpose}, \mathcal{S})} ||p' - LBS(p, W, \mathcal{J}_{tpose}, q)||,$$

where $W = \text{SKIN}(p, \mathcal{J}_{tpose})$ and (p, p') a pair of points in correspondence.

4. Experiments

We now evaluate our main contributions: that learning from long-term temporal context improves the results, that our method outperforms existing correspondence-free

	Skeletal motion				Detail preserv.		
	MPJPE (m)	PA-MPJPE (m)	Acc	PA-Acc	MPVD (m)	PA-MPVD (m)	MEDL (m)
With $\mathcal{L}_{rot}$	0.180	0.055	0.016	0.008	0.147	0.054	0.001
Without $\mathcal{L}_{rot}$	0.192	0.074	0.020	0.009	0.160	0.077	0.002

Table 2. Ablation on the rotation cycle consistency loss $\mathcal{L}_{rot}$. Using $\mathcal{L}_{rot}$ leads to a significant improvement on all metrics.

	Skeletal motion				Detail preserv.		
Context duration	MPJPE (m)	PA-MPJPE (m)	Acc	PA-Acc	MPVD (m)	PA-MPVD (m)	MEDL (m)
0.16s (5 frames)	0.195	0.081	0.014	0.009	0.161	0.074	0.001
0.33s (10 frames)	0.176	0.062	0.014	0.008	0.149	0.061	0.001
0.5s(15 frames)	0.174	0.059	0.016	0.009	0.143	0.058	0.001
1s (30 frames)	0.178	0.056	0.016	0.008	0.145	0.058	0.001
2s (60 frames)	0.194	0.061	0.016	0.008	0.160	0.059	0.001

Table 3. Evaluation of learning with different temporal contexts on SMPL test set. Including long-term context of $1s$ during training improves results overall.

methods, and that it retargets unstructured 3D human motions.

To achieve this, we first show quantitatively that considering long-term temporal context improves the accuracy of motion retargeting. Second, we present quantitative comparisons to state-of-the-art results for both geometric detail preservation and skeletal-level motion retargeting on challenging shape transfers with both naked and clothed target shapes where both shape and motions are unseen during training. Finally, we show results that retarget the raw output of a 4D multi-view acquisition platform to new target characters. More results are in supplementary material.

Data For a fair evaluation of the different methods considered in our comparison, we evaluate all methods on two test sets. The first one is a test set of representative naked human body shapes performing the same set of varying long-term motions. To build this dataset, we consider 4 body shapes created using the SMPL model [25], as is commonly done when evaluating human deformation transfer methods *e.g.* [33]. We sample body shapes at ± 2 standard deviations along the first 2 principal components to cover the main variabilities of human body shape. Skeleton-based retargeting methods *e.g.* [36] commonly evaluate on the Mixamo dataset [1]. Inspired by this, we create and retarget a set of 4 motions to all body shapes using Mixamo to generate corresponding ground truth motions. We call this *SMPL test set* in the following.

The second test set considers characters with clothing performing long-term motions. This test set allows in particular to evaluate the generalization of different methods to geometric detail in the target shape. To generate this test set, we take 4 characters with clothes from CAPE [26] and retarget 3 motions from the Mixamo dataset to each of these models. We call this *CAPE test set* in the following.

Note that for both of SMPL and CAPE test sets, none of the body shapes or motions were observed by any of the methods during training. Furthermore, by using Mixamo to retarget the same motion to different body shapes, ground truth retargeting results are available for quantitative testing.

We also propose an additional test set which considers raw untracked data acquired using a multi-view camera setup [28]. For this dataset, no ground truth retargeting is available, and we provide qualitative results for our method. We call this *multi-view test set* in the following.

Evaluation metrics The goal is to evaluate the retargeting results in terms of the overall preservation of the motion and the detail-preservation of the target geometry.

To evaluate the overall motion, we use two complementary metrics that operate exclusively on the skeletal level. First, we consider the Mean-per-joint error (MPJPE) between the ground truth and the retargeting result, which evaluates the overall accuracy of the joint positions, and the Procrustes aligned MPJPE (PA-MPJPE), which eliminates the error in global displacement. Second, to evaluate motion smoothness, we consider the mean acceleration difference between ground truth predicted motions (Acc) and its Procrustes aligned (PA-Acc) version.

To evaluate the detail-preservation of the target geometry, we use two complementary metrics that operate on the surface. The first are mean-per-vertex distance (MPVD) and Procrustes aligned MPVD (PA-MPVD) between the ground truth and the retargeting result, which evaluate the global extrinsic accuracy of the predicted surface. Second, to evaluate the preservation of intrinsic geometry, we compute a mean difference in edge length (MDEL) between the

[1] https://www.mixamo.com

ground truth and the retargeting result. As we operate on point clouds, we create edges by connecting the 6 closest neighbors of every point in the ground truth.

4.1. Learning with long-term temporal context

Our first experiment demonstrates that considering temporal context beyond a few frames during training is beneficial to motion retargeting. To this end, we train our model with motion sequences containing different numbers of frames, *i.e*. for each model, all training sequences have a fixed number of frames, which ranges from 5 frames (similar to shape deformation transfer methods [8, 33]) to 60 frames (similar to skeleton based methods [2, 22, 37]). Table 3 shows the results.

Note that including long-term context improves almost all metrics up to 15 frames, and that the overall accuracy of the retargeting improves up to 30 frames, which corresponds to $1s$ of motion. In all following experiments, we use the model trained with sequences of 30 frames.

4.2. Quantitative comparison to state-of-the-art

We now present a comparative analysis to state-of-the-art retargeting methods.

Competing methods As summarized in Table 1, there are three lines of existing methods. Skeleton-based retargeting methods are not comparable to our approach as they require hand crafted skinning weights as input, which are not available for our test sets. We therefore compare our method to a human deformation transfer method that considers short-term dynamics [33], to a method that leverages a motion prior [17] and to [39], the only correspondence-free deformation transfer method. The human deformation transfer method [33] operates on input shapes with known correspondence information, so we provide correspondences to it. For the method leveraging a human motion prior, we provide them with a full identity sequence to extract the body shape parameters required for retargeting.

Quantitative results Table 4 provides quantitative results when considering naked and clothed body shapes of the SMPL and CAPE test sets, respectively. Our method outperforms the correspondence-free motion prior [17] on almost all evaluation metrics on both datasets. In particular, the skeletal joint positions after Procrustes alignment are significantly more accurate for both test sets, and without Procrustes alignment, the mean is $4.9cm$ more accurate for naked target shapes, while being almost identical ($2mm$ worse) for clothed ones. Joint accelerations are more accurate when using our model. Geometric detail is significantly better preserved using our model when considering Procrustes alignment for both datasets, and without Procrustes alignment, the errors of both models are similar. This implies that our model retains geometric detail better, but that global alignment is not perfect. Table 4 shows that our method is significantly better than [39].

Compared to the deformation transfer method [33] that leverages correspondence information, both correspondence-free methods, [17] and our method, perform slightly worse, as expected. However, the performance of our method is close; for both test sets, all errors are within $5cm$ of Regateiro *et al*.

4.3. Animating target shape with captured 4D data

We finally demonstrate our method's performance when directly animating a target shape with the raw 4D output of a multi-view acquisition platform. To this end, we take sequences of the multi-view test set and directly retarget them to characters generated using SMPL and CAPE, respectively. Note that the input sequence suffers from acquisition noise and that no correspondence information is available, *i.e*. we input the raw untracked 4D sequence.

Fig. 3 shows the results obtained using our method. Note how the overall motion of the source sequence as well as the geometric detail of the target shape are preserved by our method. To the best of our knowledge, our method is the first that can retarget untracked 4D acquisition data online.

These examples show the robustness of our method to unseen shapes. The first source motion exhibits a body shape with hair, which was not seen during training, demonstrating the robustness of our method to unseen source shapes. The preservation of the geometry when considering the CAPE target shape also demonstrates that our model generalizes well when considering unseen target shapes.

The quantitative and qualitative results show that our method generalises well on unseen motions such us the motions from Mixamo and unseen shapes such us clothed shapes from CAPE and raw 4D output of multi-view acquisition platforms. However, we cannot generalize on clothed shapes with wide garments which is due to LBS limitations.

5. Conclusion

We proposed the first online retargeting method that allows to animate a target shape with a correspondence-free source motion. We demonstrated that including long term temporal context of $1s$ is beneficial when retargeting dense motion. Our low dimensional intermediate skeletal representation combined with the skinning field generalizes well to unseeen shapes and motions. In particular, we demonstrate that our model, learned exclusively on naked body shapes, generalizes to inputs with hair and clothing.

Interesting future works include going beyond linear blend skinning to allow for extensions to complex garments such as wide or layered clothing. One option is to explicitly include clothing in the model. Extending the solution to handle hands and expressions retargeting is also possible.

	Skeletal motion				Detail preserv.		
	MPJPE (m) ↓	PA-MPJPE (m) ↓	Acc ↓	PA-Acc ↓	MPVD (m) ↓	PA-MPVD (m) ↓	MDEL (m) ↓
Naked target shapes from SMPL test set							
Methods with correpondence information							
Deform. transfer [33]	0.152	0.028	0.005	0.004	0.130	0.028	0.001
Correpondence-free methods							
Motion prior [17]	0.238	0.096	0.019	0.014	0.152	0.078	0.402
Deform. transfer [39]	0.388	0.165	0.024	0.014	0.227	0.132	0.003
Ours	0.178	0.056	0.016	0.008	0.145	0.058	0.001
Clothed target shapes from CAPE test set							
Methods with correpondence information							
Deform. transfer [33]	0.093	0.028	0.006	0.004	0.065	0.027	0.001
Correpondence-free methods							
Motion prior [17]	0.138	0.108	0.020	0.017	0.107	0.086	0.392
Deform. transfer [39]	0.317	0.168	0.025	0.015	0.179	0.137	0.003
Ours	0.149	0.058	0.019	0.009	0.093	0.060	0.002

Table 4. Comparison to state-of-the-art on naked (top) and clothed (bottom) target shapes.

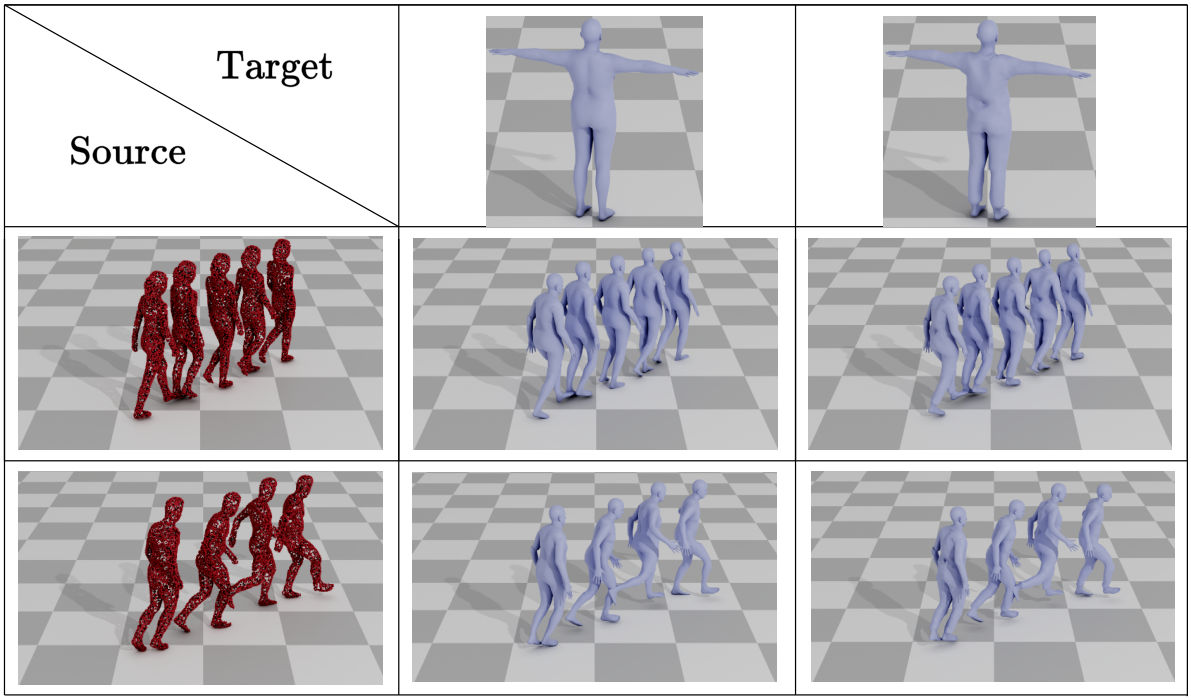

Figure 3. Animating target shapes with untracked captured 4D data directly. We consider a walking motion (top) and a kicking motion (bottom), which are retargeted to a naked (left) and clothed (right) target shape.

Potential negative societal impact

This work presents a method that allows 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.

6. Acknowledgements

We thank João Regateiro and Edmond Boyer for helpful discussions. This work was supported by French government funding managed by the National Research Agency

under the Investments for the Future program (PIA) grant ANR-21-ESRE-0030 (CONTINUUM) and 3DMOVE - 19-CE23-0013-01.

References

[1] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *Transactions on Graphics*, 39(4):62–1, 2020. 2, 3

[2] Kfir Aberman, Rundi Wu, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Learning character-agnostic motion for motion retargeting in 2d. *Transactions on Graphics*, 38(4):75:1–14, 2019. 3, 7

[3] Ilya Baran, Daniel Vlasic, Eitan Grinspun, and Jovan Popović. Semantic deformation transfer. In *SIGGRAPH*, pages 1–6. 2009. 2

[4] Jean Basset, Adnane Boukhayma, Stefanie Wuhrer, Franck Multon, and Edmond Boyer. Neural human deformation transfer. In *International Conference on 3D Vision*, pages 545–554, 2021. 2, 3

[5] Federica Bogo, Javier Romero, Matthew Loper, and Michael J Black. Faust: Dataset and evaluation for 3d mesh registration. In *Conference on Computer Vision and Pattern Recognition*, pages 3794–3801, 2014. 1

[6] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic faust: Registering human bodies in motion. In *Conference on Computer Vision and Pattern Recognition*, pages 6233–6242, 2017. 1

[7] Adnane Boukhayma, Jean-Sébastien Franco, and Edmond Boyer. Surface motion capture transfer with gaussian process regression. In *Conference on Computer Vision and Pattern Recognition*, pages 184–192, 2017. 2

[8] Haoyu Chen, Hao Tang, Nicu Sebe, and Guoying Zhao. Aniformer: Data-driven 3d animation with transformer. In *British Machine Vision Conference*, 2021. 3, 4, 7

[9] Haoyu Chen, Hao Tang, Henglin Shi, Wei Peng, Nicu Sebe, and Guoying Zhao. Intrinsic-extrinsic preserved gans for unsupervised 3d pose transfer. In *International Conference on Computer Vision*, pages 8630–8639, 2021. 3

[10] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021. 4, 5

[11] Kwang-Jin Choi and Hyeong-Seok Ko. Online motion retargetting. *The Journal of Visualization and Computer Animation*, 11(5):223–235, 2000. 2

[12] Luca Cosmo, Antonio Norelli, Oshri Halimi, Ron Kimmel, and Emanuele Rodola. Limp: Learning latent shape representations with metric preservation priors. In *European Conference on Computer Vision*, pages 19–35. Springer, 2020. 2

[13] Brian Delhaisse, Domingo Esteban, Leonel Rozo, and Darwin Caldwell. Transfer learning of shared latent spaces between robots with similar kinematic structure. In *International Joint Conference on Neural Networks*, pages 4142–4149, 2017. 2

[14] Michael Gleicher. Retargetting motion to new characters. In *SIGGRAPH*, pages 33–42, 1998. 1, 2

[15] Hanyoung Jang, Byungjun Kwon, Moonwon Yu, Seong Uk Kim, and Jongmin Kim. A variational u-net for motion retargeting. In *SIGGRAPH Asia Posters*, pages 1–2. 2018. 2

[16] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*, pages 18–35. Springer, 2020. 5

[17] Boyan Jiang, Yinda Zhang, Xingkui Wei, Xiangyang Xue, and Yanwei Fu. H4D: human 4d modeling by learning neural compositional representation. In *Conference on Computer Vision and Pattern Recognition*, pages 19355–19365, 2022. 2, 3, 5, 7, 8

[18] Pushkar Joshi, Wen C Tien, Mathieu Desbrun, and Frédéric Pighin. Learning controls for blend shape based realistic facial animation. In *Symposium of Computer Animation*, pages 17–20. 2003. 2

[19] Hyeong-Seok Ko, Kwang-Jin Choi, Min Gyu Choi, Seyoon Tak, Byoungwon Choe, and Oh-Young Song. Research problems for creating digital actors. In *Eurographics State of the Art Reports*, 2003. 2

[20] Paul G Kry, Doug L James, and Dinesh K Pai. Eigenskin: real time large deformation character skinning in hardware. In *Symposium on Computer Animation*, pages 153–159, 2002. 2

[21] Jehee Lee and Sung Yong Shin. A hierarchical approach to interactive motion editing for human-like figures. In *SIGGRAPH*, pages 39–48, 1999. 2

[22] Jongin Lim, Hyung Jin Chang, and Jin Young Choi. Pmnet: Learning of disentangled pose and movement for unsupervised motion retargeting. In *British Machine Vision Conference*, number 6:1–7, 2019. 1, 3, 7

[23] Yaron Lipman, Olga Sorkine, David Levin, and Daniel Cohen-Or. Linear rotation-invariant coordinates for meshes. *Transactions on Graphics*, 24(3):479–487, 2005. 2

[24] Lijuan Liu, Youyi Zheng, Di Tang, Yi Yuan, Changjie Fan, and Kun Zhou. Neuroskinning: Automatic skin binding for production characters with deep graph networks. *Transactions on Graphics*, 38(4):1–12, 2019. 2

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: a skinned multi-person linear model. *Transactions on Graphics*, 34(6):1–16, 2015. 4, 5, 6

[26] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 6

[27] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 1, 5

[28] Mathieu Marsot, Stefanie Wuhrer, Jean-Sebastien Franco, and Stephane Durocher. A structured latent space for human body motion generation. In *Conference on 3D Vision*, 2022. 3, 6

[29] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *Transactions On Graphics*, 39(4):82–1, 2020. 5

[30] Albert Mosella-Montoro and Javier Ruiz-Hidalgo. Skinningnet: Two-stream graph convolutional neural network for skinning prediction of synthetic characters. In *Conference on Computer Vision and Pattern Recognition*, pages 18593–18602, 2022. 2, 4

[31] Xuming Ouyang and Cunguang Feng. Autoskin: Skeleton-based human skinning with deep neural networks. In *Journal of Physics: Conference Series*, volume 1550, page 032163. IOP Publishing, 2020. 4

[32] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 4, 11

[33] João Regateiro and Edmond Boyer. Temporal shape transfer network for 3d human motion. In *International Conference on 3D Vision*, 2022. 2, 3, 4, 6, 7, 8

[34] Robert W Sumner and Jovan Popović. Deformation transfer for triangle meshes. *Transactions on Graphics*, 23(3):399–405, 2004. 2

[35] Seyoon Tak and Hyeong-Seok Ko. A physically-based motion retargeting filter. *Transactions on Graphics*, 24(1):98–117, 2005. 2

[36] Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, and Jun Saito. Contact-aware retargeting of skinned motion. In *International Conference on Computer Vision*, 2021. 2, 3, 6

[37] 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, 2, 3, 4, 5, 7, 11

[38] Timo Von Marcard, Roberto Henschel, Michael J 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*, pages 601–617, 2018. 1

[39] Jiashun Wang, Chao Wen, Yanwei Fu, Haitao Lin, Tianyun Zou, Xiangyang Xue, and Yinda Zhang. Neural pose transfer by spatially adaptive instance normalization. In *Conference on Computer Vision and Pattern Recognition*, pages 5831–5839, 2020. 2, 3, 7, 8

[40] Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Landreth, and Karan Singh. Rignet: Neural rigging for articulated characters. *Transactions on Graphics*, 39(4):58:1–14, 2020. 2, 4

[41] Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *Conference on Computer Vision and Pattern Recognition*, pages 13284–13293, 2021. 2, 4

[42] Yizhou Yu, Kun Zhou, Dong Xu, Xiaohan Shi, Hujun Bao, Baining Guo, and Heung-Yeung Shum. Mesh editing with poisson-based gradient field manipulation. In *SIGGRAPH*, pages 644–651, 2004. 2

[43] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *International Conference on Computer Vision*, pages 7114–7123, 2019.

[44] Keyang Zhou, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Unsupervised shape and pose disentanglement for 3d meshes. In *European Conference on Computer Vision*, pages 341–357, 2020. 2

[45] Kun Zhou, Weiwei Xu, Yiying Tong, and Mathieu Desbrun. Deformation transfer to multi-component objects. In *Computer Graphics Forum*, volume 29, pages 319–325, 2010. 2

Supplementary material

In this supplementary material, we provide the network architecture of our framework and implementation details.

Our framework consists of three components: skeleton regressor SKR, skeletal motion retargeting SMRM, and skinning predictor SKIN. The following provides details for each of them. 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 [**?**] with a learning rate of 0.0001.

The training of the whole framework takes 18 hours on GeForce RTX 2080 TI. The whole model has 7.6 million trainable parameters. The skeleton regressor has 1.6M parameters, the locomotion retargeting module has 5.8M parameters and the skinning predictor has 0.2M parameters.

Skeleton Regressor We detail the architecture of this network in Table 5. The feature transformation STNkd is a network inspired by PointNet [32] in order to make the point cloud $\mathcal{S}$ rotation invariant, k is the feature size of the input to STNkd.

Index	Inputs	Operation	Output shape	Activation
(1)	Input	Source scan $\mathcal{S}^A$	$N \times 3 \times V$	-
(1')	(1)	Feature-Transformation STNkd ($k = 3$)	$N \times 3 \times V$	-
(2)	(1')	conv1d ($3 \rightarrow 64, 1 \times 1$)	$N \times 64 \times V$	Relu
(2')	(2)	Feature-Transformation STNkd ($k = 64$)	$N \times 64 \times V$	-
(3)	(2')	conv1d ($64 \rightarrow 128, 1 \times 1$)	$N \times 128 \times V$	Relu
(4)	(3)	conv1d ($128 \rightarrow 1024, 1 \times 1$)	$N \times 1024 \times V$	Relu
(5)	(4)	Maxpooling	$N \times 1024$	-
(6)	(5)	Linear ($1024 \rightarrow 512$)	$N \times 512$	Relu
(7)	(6)	Linear ($512 \rightarrow 256$) Dropout ($p = 0.3$)	$N \times 256$	Relu
(8)	(7)	Linear ($256 \rightarrow J \times 3$)	$N \times 3 \times J$	-

Table 5. The network architecture for skeleton regressor.

The network is trained on 100 epochs where each epoch sees 63520 examples.

Skeletal motion retargeting The SMRM model is a reimplementation of [37]. We summarize the architecture in Table 6. The FK is a forward kinematic layer.

The network is trained on 1000 epochs where each epoch sees 2330 examples.

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 ($p = 0.2$)	$N \times 512 \times J$
(3)	(2)	Linear ($512 \rightarrow 4 \times J$)	$N \times 4 \times J$
(4)	(3)	Linear ($512 \rightarrow 3$)	$N \times 3$
(5)	(3)+ (1')	FK	$N \times 3 \times J$

Table 6. The network architecture for Skeletal motion retargeting.

Skinning predictor We detail the architecture of the skinning predictor in Table 7.

Index	Inputs	Operation	Output shape	Activation
(1)	$N \times 3 \times J$	Linear ($J \times 3 \rightarrow 256$) 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 7. The network architecture for Skinning predictor.

The network is trained on 200 epochs where each epoch sees 63520 examples.