Correspondence-free online human motion retargeting

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

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

Abstract

We present a data-driven framework for unsupervised human motion retargeting that animates a target subject with the motion of a source subject. Our method is correspondence-free, requiring neither spatial correspondences between the source and target shapes nor temporal correspondences between different frames of the source motion. This allows to animate a target shape with arbitrary sequences of humans in motion, possibly captured using 4D acquisition platforms or consumer devices. Our method unifies the advantages of two existing lines of work, namely skeletal motion retargeting, which leverages long-term temporal context, and surface-based retargeting, which preserves surface details, by combining a geometry-aware deformation model with a skeleton-aware motion transfer approach. This allows to take into account long-term temporal context while accounting for surface details. During inference, our method runs online, i.e. *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 for skeletal motion and detail preservation. Furthermore, our method generalizes to unobserved motions and body shapes. We demonstrate that our method achieves state-of-the-art results on two test datasets and that it can be used to animate human models with the output of a multi-view acquisition platform. Code is available at* https://gitlab.inria.fr/rrekikdi/human-motion-retargeting2023.

1. Introduction

Human motion retargeting is the process of animating a target character with the motion of a source character. We study this problem for densely sampled 3D surfaces, and arbitrary motion duration. This has applications in video gaming, movie making, avatar animation, and augmenting existing datasets of 4D body motions *e.g.* [6, 7, 28, 39].

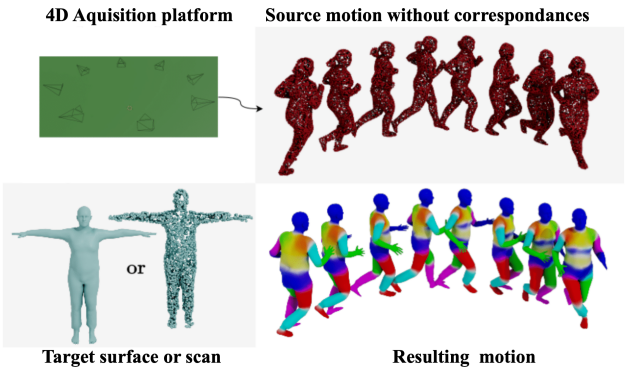

Figure 1. Given an untracked motion performed by a source subject (top) and a target body shape (bottom left), our method animates the target with the source motion, preserving temporal correspondences of the output motion (bottom right).

Motion retargeting has recently made significant progress. Skeleton-based methods can take long-term temporal context of about 2 seconds into account *e.g.* [15, 23, 37]. Surface deformation-based methods allow to retarget geometric details while possibly capturing short time dynamics *e.g.* [5, 34, 35, 40].

Most of these works take as input structured data such as skeletons or template-aligned surfaces for which correspondences are known.

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. Solving the retargeting problem for unstructured 3D data is challenging as computing correspondences is a combinatorial problem. Fig. 1 illustrates the problem's input and output.

[†] Institute of Engineering Univ. Grenoble Alpes

Recently, a first solution for correspondence-free motion retargeting has been proposed [18], which 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. Our method is able to learn temporal context of $1s$ duration, which we demonstrate to improve the accuracy of motion retargeting on challenging examples.

We achieve this by combining the advantages of skeletal and surface-based motion retargeting to learn both temporal context and geometric detail. More specifically, we combine skeleton-based motion retargeting and shape deformation modeling, which allows 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 recent methods to deform the shape based on skeletal deformations [25, 31, 41, 42] can accurately model geometric detail. Our hypothesis is that the locomotion information of the source sequence is explained at the skeletal level, while the dense surface details are intrinsic to the target shape. Thus, our idea is to minimize a motion preservation loss that transfers the skeletal motion of the source to the target, and a shape preservation loss that keeps the shape details of the target.

Our main contributions are:

- The first correspondence-free online approach for dense human motion retargeting that learns temporal context.
- A demonstration that long-term temporal context improves the accuracy of motion retargeting.
- An approach that outperforms correspondence-free methods in both geometric detail preservation and skeletal-level motion retargeting.

2. Related Work

Works addressing motion retargeting can be categorized into three groups: works operating at the skeleton level, works transferring shape deformations at a dense geometric level, and motion priors 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*. [15] 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 [24, 43] and using blending techniques [19, 21].

Follow up works [12, 20, 22, 36] 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 [2, 14, 16, 37] have shown outstanding results without requiring many handcrafted energies. Some of these data-driven approaches [14, 16] require paired training data, whose design involves human effort. Therefore, another line of works [2, 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 [38] 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.

Similar to our approach, a recent method combines skeletal-level and geometric information for motion retargeting [44]. Unlike our work, this method focuses on contact preservation, in particular for characters created with computer aided design tools for which clean skeletal models are provided as input. In contrast, we focus on retargeting the raw output from acquisition platforms directly, without the need for costly and error-prone pre-processing to find correspondence information or fit skeletons.

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 [13, 48] or by optimising mesh deformations [4, 8, 35, 47]. Some works [4, 35] 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 [5] consider motion retargeting as shape deformation transfer, encoding the shape identity of the target character and transferring it to the source character. A recent work adds short-term temporal context by considering 4 consecutive frames [34]. These works showed impressive results by operating on meshes in correspondence. As providing spatio-temporal correspondences between source and target meshes is a difficult task, Wang *et al*. [40] 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. Chen *et al*. [9] extend this work to include short-term temporal information of 3 frames. This improves results at the cost of requiring a target sequence instead of a single

target frame as input during training. 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 outperforms neural pose transfer (NPT) [40], the only existing correspondence-free method that can be trained without sequence-level supervision.

2.3. Generic motion priors

Generic motion priors representing dense human motion in a latent space that disentangles locomotion and dense body shape information have recently been applied to motion retargeting [18, 29]. They allow to retarget the body motion to different body shape while considering long-term temporal context. Marsot *et al*. [29] use structured data and learn correlations between dense body shape and skeletal locomotion. Closer to our setting, H4D [18] considers unstructured point clouds and retains dense surface details of the target shape. While motion priors account for long-term temporal context, they cannot easily be applied to sequences of arbitrary duration at inference as they operate on normalized temporal segments of motion. We demonstrate experimentally that our method outperforms H4D [18].

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 shape deformation models, we propose the first correspondence-free retargeting approach that learns long-term temporal context, allows for arbitrary duration at inference, and models geometric detail.

Method	Temporal context	Online inference	Geometric detail	Unstructured
Skeleton-based				
[1, 2, 23, 37]	✓	✓	✗	✗
[38, 44]	✓	✓	✓	✗
Shape deformation transfer				
[5, 10, 34]	✗	✓	✓	✗
[9, 40]	✗	✓	✓	✓
Motion priors				
[29]	✓	✗	✓	✗
[18]	✓	✗	✓	✓
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.

3. Motion retargeting method

Our motion retargeting method takes as input a source motion of subject A given as sequence of n point clouds $\{\mathcal{S}_i^A\}_{i=1}^n$ without temporal correspondences, and the target shape B in T-pose given as point cloud, possibly with connectivity information, $\mathcal{S}_{tpose}^B$. Our objective is to generate a sequence of retargeted scans $\{\mathcal{S}_i^B\}_{i=1}^n$, where $\mathcal{S}_i^B$ imitates the pose of $\mathcal{S}_i^A$ while retaining the body shape of $\mathcal{S}_{tpose}^B$, and is in correspondence with $\mathcal{S}_{tpose}^B$.

A common strategy for motion retargeting is to disentangle the high-level motion from the shape deformation caused by body shape, and to combine the source motion with the target body shape. This is computationally expensive when considering densely sampled input surfaces containing thousands of vertices per frame, especially if the surfaces are not in correspondence, which hinders existing approaches to learn long-term temporal context.

We overcome this challenge by making the hypothesis that source motion information is fully explained at a skeletal level, while geometric detail information is encoded in the target. This hypothesis leads to a retargeting objective consisting of two losses to be optimized during training. The first loss called source motion preservation loss $\mathcal{L}_{motion}$ encourages the retargeted motion to resemble $\{\mathcal{S}_i^A\}_{i=1}^n$ at a skeletal level. Considering $\mathcal{L}_{motion}$ on a skeletal level 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. This allows to train using long-term context without running into complexity issues. The second loss called target geometry preservation loss $\mathcal{L}_{geom}$ encourages each frame of the output sequence to have similar geometric details as $\mathcal{S}_{tpose}^B$. Considering $\mathcal{L}_{geom}$ independently per-frame allows processing high-dimensional point clouds that contain significant geometric detail. Our training procedure aims to optimize

$$\mathcal{L} = \mathcal{L}_{motion} + \mathcal{L}_{geom}. \tag{1}$$

During inference, our method then retargets the motion of A to B at the skeletal level and uses the resulting skeletal motion of B to transfer geometric details from $\mathcal{S}_{tpose}^B$ in a forward pass. This allows for efficient online retargeting. Fig. 2 gives a visual overview of our method.

3.1. Preserving source motion

Our method aims to preserve the motion of the source character at a skeletal level. Working on the skeletal level is a key ingredient of our work, as this reduced representation allows to train using long-term temporal context without

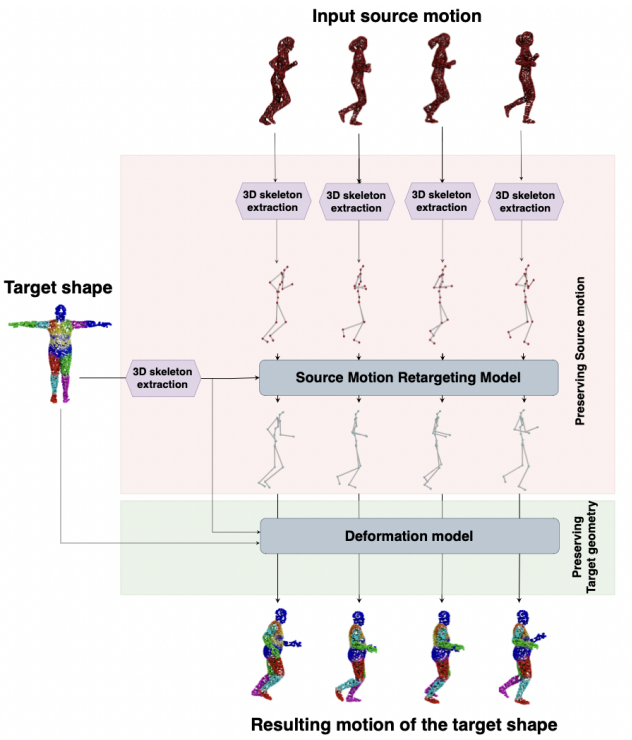

Figure 2. Our method takes a source sequence of unstructured point clouds and a target point cloud as input, and outputs the target character performing the input motion. The method first retargets the source motion to the target character at the skeletal level (red part), and then adds the target's surface details to the resulting motion using a deformation model (green part).

running into complexity issues. We first detail the architecture of our network and define the source motion preservation loss $\mathcal{L}_{motion}$, then outline how a skeletal representation is computed based on the inputs $\{\mathcal{S}_i^A\}_{i=1}^n$ and $\mathcal{S}_{tpose}^B$ of our method. In the following, $\{\mathcal{J}_i^A\}_{i=1}^n$ and $\mathcal{J}_{tpose}^B$ denote skeletal representations of $\{\mathcal{S}_i^A\}_{i=1}^n$ and $\mathcal{S}_{tpose}^B$.

Architecture We consider the task of transferring the skeletal motion $\{\mathcal{J}^A\}_{i=1}^n$ to $\mathcal{J}_{tpose}^B$. Due to the lack of motion captures containing two different subjects performing the same motion, we train in an unsupervised way without using ground truth pairings between a motion and its retargeted version. To solve this problem, we were inspired by a skeletal motion retargeting model (SMRM) [37] which given $\{\mathcal{J}_i^A\}_{i=1}^n$ and $\mathcal{J}_{tpose}^B$ outputs the relative joint rotations $\{\theta_i^B\}_{i=1}^n$.

SMRM starts by extracting high level locomotion features from $\{\mathcal{J}_i^A\}_{i=1}^n$ using a first recurrent network. These features are leveraged along with $\mathcal{J}_{tpose}^B$ to generate joint rotations $\{\theta_i^B\}_{i=1}^n$ using a second recurrent unit. The resulting rotations are applied to $\mathcal{J}_{tpose}^B$ using a differentiable forward kinematics layer to generate the retargeted skeletal motion $\{\mathcal{J}_i^B\}_{i=1}^n$.

Definition of $\mathcal{L}_{motion}$ To train our network, we use cycle consistency losses, an adversarial loss, and a loss encouraging temporally smooth motions, all evaluated at the skeletal level. The cycle consistency losses transfer the motion from subject A to B, and the resulting motion back to A, and compare the resulting motion of A with its initial motion. Our loss is

$$\begin{aligned}\mathcal{L}_{motion} &= \mathcal{L}_{cycon} + \mathcal{L}_{adv} \\ &+ \lambda_{rot}\mathcal{L}_{rot} + \lambda_{smooth}\mathcal{L}_{smooth} \end{aligned} \quad (2)$$

with weights $\lambda_{rot} = 0.01$ and $\lambda_{smooth} = 0.1$.

This loss is inspired by SMRM, which uses cycle consistency loss $\mathcal{L}_{cycon} = \text{MSE}(\hat{\mathcal{J}}_i^A, \mathcal{J}_i^A)$, where MSE is the mean squared error, on joint positions and adversarial loss $\mathcal{L}_{adv}$. $\mathcal{L}_{adv}$ leverages a discriminator network which is trained in a min-max game with the retargeting network to differentiate between real and retargeted motions. To improve performance, we add two additional loss functions to regularize the generated motions. Our first loss is a cycle consistency loss on the joint rotations, which allows to prevent unrealistic motions, and can be written as $\mathcal{L}_{rot} = \text{MSE}(\hat{\theta}_i^A, \theta_i^A)$. Rather than representing joint rotations using quaternions, we represent them in 6D as this representation was shown to be beneficial for deep learning applications [49]. The second loss allows for temporal smoothing of the motions by aiming to reconstruct velocities, and can be written as $\mathcal{L}_{smooth} = \text{MSE}(\Delta\hat{\mathcal{J}}_i^A, \Delta\mathcal{J}_i^A)$, with $\Delta\mathcal{J}_i = \mathcal{J}_{i+1} - \mathcal{J}_i$.

3D skeleton extraction It remains to outline how skeletons $\{\mathcal{J}^A\}_{i=1}^n$ and $\mathcal{J}_{tpose}^B$ are extracted from $\{\mathcal{S}_i^A\}_{i=1}^n$ and $\mathcal{S}_{tpose}^B$. To allow the processing of arbitrarily long input sequences, the skeleton extraction operates per-frame. To operate on unstructured point clouds, the skeleton extraction needs to be robust w.r.t. the number and order of the observed points. To achieve this, we use the PointFormer [46] architecture to extract order-invariant features from the point cloud, followed by a multi layer perceptron to regress the joint positions from these features. We choose to use the PointFormer architecture as it extracts local features of the point cloud allowing for precise joint predictions and good generalization to unseen poses. We denote this function, which recovers parameterized skeletons by $\mathcal{J} = \text{skeleton}(\mathcal{S})$. The skeleton extraction is trained in a supervised way while minimizing the MSE between the ground truth joints $\mathcal{J}$ and the predicted ones as $\mathcal{L}_{\text{skeleton}} = \text{MSE}(\text{skeleton}(\mathcal{S}), \mathcal{J})$. Our method is independent of the skeleton parameterization that is chosen. In our implementation, the skeleton is parameterized by 22 joint positions corresponding to all joints except the root joint of the SMPL body model [26].

3.2. Preserving target geometry

Our method aims to preserve the detailed geometry of the target character at a surface level. We achieve this by learning a spatially implicit geometry aware deformation model allowing to repose target character $\mathcal{S}^B_{tpose}$ into an arbitrary pose θ^B defined on skeleton $\mathcal{J}^B_{tpose}$. Defining an implicit model offers the key advantage of allowing for arbitrarily sampled input during inference. We outline the deformation model then detail our network architecture and define the target geometry preservation loss $\mathcal{L}_{geom}$.

Deformation model A common strategy to animate 3D meshes 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. While recent works in this area show impressive results [11, 31, 32, 41, 42], they are typically applicable to registered data or supervised with skinning weights during training, which are not given in our case.

To allow for arbitrarily sampled input point clouds, we learn a geometry-aware deformation model that can be evaluated for any surface point of a human in T-pose. Inspired by works that build human shape spaces [3, 26], we model deformations using two parts: a skinning weights predictor and a pose-corrective offset predictor. The pose corrective offset is added to vertices in T-pose before applying linear blend skinning to diminish artifacts. We choose this model for its simplicity and for its excellent performance when modeling human deformations.

Given the target joints $\mathcal{J}^B_{tpose}$, the set of rotations $\{\theta^B_i\}^n_{i=1}$, the skinning weights W^B_j and the pose corrective offsets $\{o^B_{i,j}\}^n_{i=1}$ of point p_j of $\mathcal{S}^B_{tpose}$, we denote the function to generate the j-th vertex of frame i by

$$\mathcal{S}^B_{i,j} = \text{retarget}\left(\theta^B_i, \mathcal{J}^B_{tpose}, W^B_j, p_j + o^B_{i,j}\right). \quad (3)$$

Architecture Our architecture consists of two networks. First, a skinning weights predictor network W_{net} that takes the target joints $\mathcal{J}^B_{tpose}$ and a 3D point p of $\mathcal{S}^B_{tpose}$ as input and outputs one skinning weight per joint $W = W_{net}(\mathcal{J}^B_{tpose}, p)$. Inspired by [11], we constrain the predicted skinning weights to sum to 1 using a softmax activation. Second, a pose-corrective offsets network O_{net} takes relative joint rotations θ^B and p as input and outputs an offset vector in $\mathbb{R}^3$ as $o^B = O_{net}(\theta^B, p)$. The networks are modeled by two three layer MLPs. The first MLP models W_{net} and is followed by a softmax activation. The second MLP models O_{net}.

Definition of $\mathcal{L}_{geom}$ Our networks are trained on pairs of frames of the same identity with known poses $\mathcal{S}_{tpose}$ and $\mathcal{S}_\theta$ with joint rotations θ. This allows for a point to point reconstruction loss. Consider point p of $\mathcal{S}_{tpose}$ and its corresponding point p' of $\mathcal{S}_\theta$. Our loss can be written as

$$\mathcal{L}_{geom} = \sum_{(p,p')} ||p' - \text{retarget}(\theta, \mathcal{J}_{tpose}, W, p + o)||^2, \quad (4)$$

where $W = W_{net}(\mathcal{J}_{tpose}, p)$ and $o = O_{net}(\theta, p)$.

3.3. Training

Inspired by recent works [17, 18, 30, 45], we choose a stage-wise training strategy to improve stability and reduce computational complexity. This is done by optimizing for $\mathcal{L}_{motion}$ and $\mathcal{L}_{geom}$ in two independent stages using skeletal motions for $\mathcal{L}_{motion}$ and densely sampled static point clouds for $\mathcal{L}_{geom}$.

For training, we leverage the AMASS dataset [28], which contains a collection of motion capture datasets that have been fitted by a parametric body model [26] to obtain dense per-frame representations. For each frame, aligned surface and skeleton information is given. All the motion sequences are temporally aligned at 30 frames per second (FPS). As training data, we consider a subset of 120 body shapes performing 2536 different motions for a total of 65000 seconds of motion. As validation data, we consider a subset of 3 body shapes performing 24 different motions for a total of 147 seconds of motion and as test set we consider 7 body shapes performing 44 motions for a total of 1715 second of motion. We call this test set *AMASS test set*.

All networks handling geometrically detailed data are trained by sampling one frame for each second of motion. To avoid learning the topology bias of a template mesh, we randomly uniformly sample N points on the surface of each 3D mesh and add Gaussian noise to generate input scans.

To train the motion retargeting at the skeletal level, we randomly sample sequences of fixed duration from AMASS and consider their skeletal motion only. The 3D skeleton extractor is trained using static 3D point clouds sampled from AMASS with their corresponding skeletons.

The geometric deformation model is trained using static AMASS data by considering pairs of frames $\{\mathcal{S}_{tpose}, \mathcal{S}_\theta\}$ of a same person in correspondence and with known rotations. To preserve correspondences while removing the bias due to SMPL topology, the models are uniformly resampled while preserving correspondence information.

4. Experiments

We now evaluate our main contributions: that learning from long-term temporal context improves the results, that our method outperforms state-of-the-art, and generalizes to previously unseen target shapes and source motions. We first show quantitatively that considering long-term temporal context improves the accuracy of motion retargeting. Second, we present quantitative comparisons using both stan-

dard error measures and evaluations via user studies to state-of-the-art results, considering both geometric detail preservation and skeletal-level motion retargeting. We consider challenging shape transfers with both naked and clothed target shapes where both shapes 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.

Data For a fair evaluation of the different methods considered in our comparison, we evaluate all methods on two test sets. First, a test set of representative naked human body shapes performing the same set of long-term motions. To build this dataset, we consider 4 body shapes created using the SMPL model [26], as is commonly done when evaluating human deformation transfer methods *e.g.* [34]. We sample body shapes at ± 2 standard deviations along the first 2 principal components to cover the main variabilities of human body shape, namely female, male, skinny and corpulent. Skeleton-based retargeting methods *e.g.* [38] commonly evaluate on the Mixamo dataset [‡]. 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*. The second test set considers characters with clothing performing long-term motions. This test set allows to evaluate the generalization of different methods to geometric detail in the target shape. To generate this test set, we take 4 characters with tight clothes from CAPE [27], namely using the outfits called "short long", "short short", "long short" and "long long" in CAPE, and retarget 4 motions from the Mixamo dataset to each of these models. We call this *CAPE test set*. Note that for both of SMPL and CAPE test sets, the 4 motions, namely "Walking", "Jogging", "Throw" and "Hip Hop Dancing", are varied in terms of global trajectory and local motions and 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 use a test set of untracked data acquired using a multi-view camera setup [29]. For this dataset, no ground truth retargeting is available, and we provide qualitative results. We call this *multi-view test set*.

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 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 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 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 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. 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 [9, 34]) to 60 frames (similar to skeleton based methods [1, 23, 37]). Table 2 shows the results. Including long-term context improves almost all metrics up to 30 frames. 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 motion retargeting methods.

Competing methods As summarized in Table 1, there are three lines of existing methods. Skeleton-based 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 correspondence-free deformation transfer methods and motion priors. For deformation transfer methods, we compare

	Skeletal motion				Detail preserv.		
Context duration	MPJPE (m)↓	PA-MPJPE (m)↓	Acc↓	PA-Acc↓	MPVD (m)↓	PA-MPVD (m)↓	MEDL (mm)↓
0.16s (5 frames)	0.203	0.073	0.021	0.009	0.158	0.061	**0.505**
0.33s (10 frames)	0.167	0.050	0.021	0.008	0.137	0.045	0.507
0.5s(15 frames)	0.161	0.046	**0.020**	0.008	0.136	0.044	0.519
1s (30 frames)	**0.158**	**0.043**	0.021	**0.008**	0.134	**0.040**	0.524
2s (60 frames)	0.160	0.044	0.023	0.008	**0.131**	0.041	0.509

Table 2. Learning with different temporal contexts on SMPL test set. Training with long-term context of $1s$ improves the results.

[‡]https://www.mixamo.com

	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							
H4D [18]	0.238	0.096	**0.019**	0.014	0.152	0.078	402.238
NPT [40]	0.388	0.165	0.024	0.014	0.227	0.132	2.739
Ours	**0.158**	**0.043**	0.021	**0.008**	**0.134**	**0.040**	**0.524**
Clothed target shapes from CAPE test set							
H4D [18]	0.161	0.091	**0.019**	0.014	0.096	0.074	395.683
NPT [40]	0.373	0.168	0.022	0.013	0.173	0.135	2.845
Ours	**0.105**	**0.040**	0.023	**0.008**	**0.075**	**0.038**	**0.539**

Table 3. Comparison to state-of-the-art on naked (top) and clothed (bottom) target shapes. Best performing scores shown in bold.

to NPT [40]. Aniformers [9] requires ground truth pairings of different individuals performing the same motion during training, which are not available in our case, making comparison impossible. For motion priors, we compare to H4D [18]. We provide a full identity sequence to this method to extract body shape parameters.

Quantitative results Table 3 provides quantitative results when considering shapes of the SMPL and CAPE test sets. Our method significantly outperforms NPT on almost all metrics on both datasets. The reason is that NPT operates per frame without any temporal context and sometimes leads to results that are temporally inconsistent. Our method also outperforms H4D 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 within $2mm$ 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.

Comparison to retargeting method using correspondences As we outperform correspondence free retargeting methods by a large margin, we also compare our method to TST [34], a state-of-the-art deformation transfer method considering short-term temporal dynamics that requires point-to-point correspondences at training and inference, and provide this method with correspondences. Our method only performs slightly worse than TST on the evaluation metrics, with all errors being within $2cm$ of TST. This performance, close to a method leveraging correspondences, highlights the potential of our correspondence-free method. Detailed numerical results of evaluation metrics are provided in supplementary material.

4.3. User Study

We further evaluate how our retargeting results are perceived by humans w.r.t. both preservation of motion and geometric detail by conducting a user study.

Design We designed a within-subject perceptual experiment with retargeting methods as a main factor. Since NPT performs significantly worse than competing methods quantitatively, we did not include it in the perceptual evaluation in order to minimize the duration of the experiment for each user. We compared the results of our method to H4D and TST, which uses correspondences. 30 unpaid volunteers participated in this study. For each trial, they were presented stimuli showing a source motion, a target shape, and a retargeting result and were asked to rate motion and shape similarity respectively with source motion and target shape on a 7-point Likert scale. 15 repetitions were performed for each method using 3 types of motion (walking, hip-hop, throwing) and 5 different shape transfers. Participants performed 45 trials, presented in a randomized order. Statistical analysis was done using ART-ANOVAs and post-hoc Tukey tests.

Results Results highlight a main effect of the method both on motion and shape components ($p<0.0001$). Post-hoc analysis shows that users rated our method to preserve motion and body shape significantly better than H4D ($p<0.0001$). TST outperforms our method on motion preservation ($p<0.001$), while our method better preserves the body shape of the target ($p<0.0001$). More details about the design and results are provided in supplementary.

4.4. Ablations

We provide ablations for the architectures optimizing $\mathcal{L}_{motion}$ and $\mathcal{L}_{geom}$.

Source motion preservation We first evaluate motion transfer on the skeletal level on the densely sampled geometry of the SMPL test set. Table 4 shows that using the 6D rotation representation and the rotation supervision using $\mathcal{L}_{rot}$ improve the retargeting. The transition from quaternion to 6D rotation leads to a significant improvement on almost all metrics. Supervising with $\mathcal{L}_{rot}$ also leads to an improvement in the geometric detail preservation around the feet, wrists and head joints. This is because these joints are leaves of the hierarchical skeleton, so they have no influence on $\mathcal{L}_{cycon}$ but they do influence $\mathcal{L}_{rot}$.

To evaluate the 3D skeleton extraction, we use the AMASS test set. We compare between PointNet [33], which considers global features and PointFormer [46], which introduces local features. Using PointFormer improves the MPJPE to **21mm** over **46mm** for PointNet.

Target geometry preservation To ablate the deformation model used to optimize $\mathcal{L}_{geom}$, we use the AMASS test set. Using pose-corrective offsets improves precision by reducing the average reconstruction error from **8.4mm** to **5mm**. Further ablations are in supplementary.

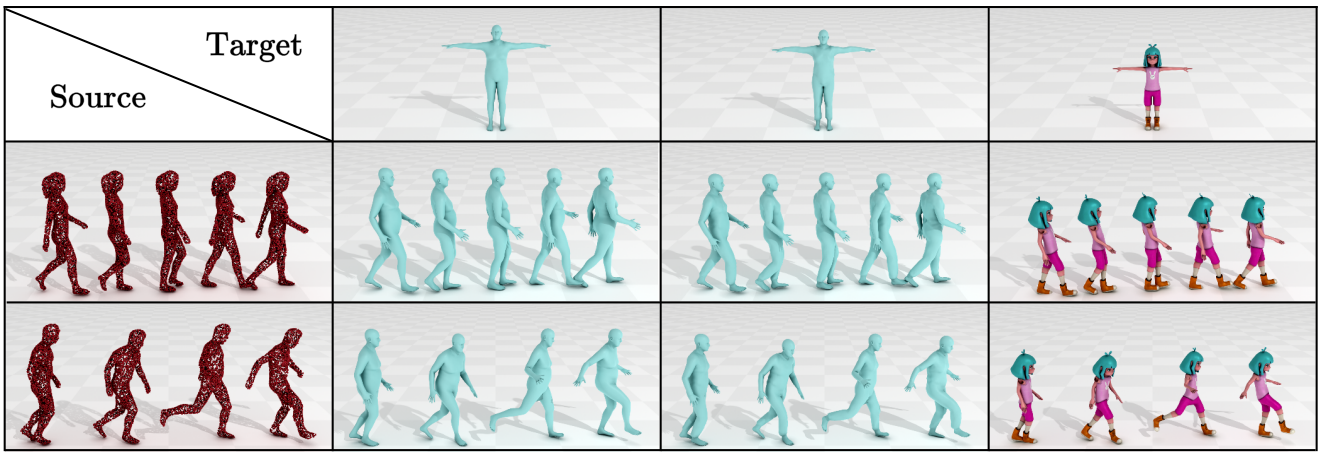

Figure 3. Animating target shapes with untracked captured 4D data directly. We consider a walking motion (top) and a kicking motion (bottom), which are retargeted to a naked (left),clothed (middle) and a CAD-generated (right) target shape.

	Skeletal motion				Detail preserv.		
	MPJPE (m) ↓	PA-MPJPE (m) ↓	Acc ↓	PA-Acc ↓	MPVD (m) ↓	PA-MPVD (m) ↓	MEDL (m) ↓
[37]	0.163	0.056	0.023	0.011	0.165	0.089	1.597
[37]+6D	0.159	0.043	0.024	0.008	0.137	0.048	0.684
[37]+$\mathcal{L}_{rot}$	0.161	0.044	0.021	0.008	0.156	0.076	1.445
[37]+$\mathcal{L}_{rot}$+6D	**0.158**	**0.043**	**0.021**	**0.008**	**0.134**	**0.040**	**0.524**

Table 4. Ablation of motion retargeting. Quantitative evaluation on SMPL shapes with different rotation representation for θ and ablation of the rotation cycle consistency loss $\mathcal{L}_{rot}$.

4.5. Animating target shape with captured 4D data

We demonstrate our method's performance when animating a target shape with the raw 4D output of a multi-view acquisition platform. We retarget sequences of the multi-view test set directly to characters generated using SMPL, CAPE, and a Mixamo character designed using computer aided design (CAD) tools. Note that the input sequence suffers from acquisition noise and that no correspondence information is available, *i.e*. we input the untracked 4D sequence.

Fig. 3 shows the results obtained using our method. Note how the motion of the source sequence as well as the geometric detail of the different target shapes 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, not seen during training, demonstrating the robustness of our method to unseen source shapes. The preservation of the geometry for CAPE and CAD generated target shapes also demonstrates that our model generalizes well when considering unseen target shapes.

Quantitative and qualitative results show our method to generalise well on unseen motions *e.g*. from Mixamo and unseen shapes *e.g*. clothed shapes from CAPE and untracked 4D output of multi-view acquisition platforms.

4.6. Limitations

While our method gives state-of-the-art results for the correspondence-free motion retargeting problem, limitations remain. It cannot generalize on clothed shapes with wide garments due to limitations resulting from our deformation model. The method is also restricted to shapes that can be parameterized by a skeleton with fixed topology and is not able to capture detailed hand or facial motion. Potential negative societal impact is discussed in supplementary.

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 prior 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.

Future works include allowing 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 hands and expressions is also interesting.

6. Acknowledgement

We thank João Regateiro and Edmond Boyer for helpful discussions. This work was funded by the French National Research Agency (ANR) 3DMOVE - 19-CE23-0013.

References

[1] 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, 6

[2] 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

[3] Brett Allen, Brian Curless, Zoran Popović, and Aaron Hertzmann. Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 147–156. Citeseer, 2006. 5

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

[5] 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. 1, 2, 3

[6] 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

[7] 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

[8] 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

[9] Haoyu Chen, Hao Tang, Nicu Sebe, and Guoying Zhao. Aniformer: Data-driven 3d animation with transformer. In *British Machine Vision Conference*, 2021. 2, 3, 6, 7

[10] 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

[11] 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. 5

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

[13] 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

[14] 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

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

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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

[21] 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

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

[23] 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, 6

[24] 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

[25] 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

[26] 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

[27] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 6

[28] 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

[29] Mathieu Marsot, Stefanie Wuhrer, Jean-Sebastien Franco, and Stephane Durocher. A structured latent space for hu-

man body motion generation. In *Conference on 3D Vision*, 2022. 3, 6

[30] 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

[31] 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, 5

[32] Xuming Ouyang and Cunguang Feng. Autoskin: Skeleton-based human skinning with deep neural networks. In *Journal of Physics: Conference Series*, page 032163. IOP Publishing, 2020. 5

[33] 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. 7

[34] João Regateiro and Edmond Boyer. Temporal shape transfer network for 3d human motion. In *International Conference on 3D Vision*, 2022. 1, 2, 3, 6, 7

[35] Robert W Sumner and Jovan Popović. Deformation transfer for triangle meshes. *Transactions on Graphics*, 23(3):399–405, 2004. 1, 2

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

[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, 6, 8

[38] 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

[39] 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

[40] 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. 1, 2, 3, 7

[41] 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, 5

[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13284–13293, 2021. 2, 5

[43] 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

[44] Jiaxu Zhang, Junwu Weng, Di Kang, Fang Zhao, Shaoli Huang, Xuefei Zhe, Linchao Bao, Ying Shan, Jue Wang, and Zhigang Tu. Skinned motion retargeting with residual perception of motion semantics & geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13864–13872, 2023. 2, 3

[45] 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. 5

[46] 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. 4, 7

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

[48] 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

[49] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 4