Learning Nonlinear Soft-Tissue Dynamics for Interactive Avatars

DAN CASAS, Universidad Rey Juan Carlos, Madrid
MIGUEL A. OTADUY, Universidad Rey Juan Carlos, Madrid

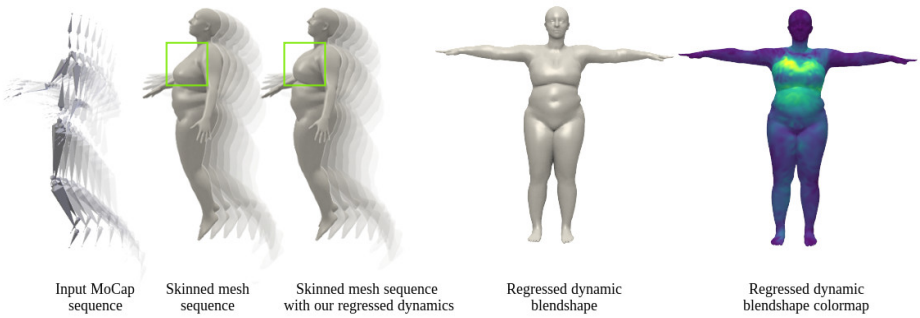


Fig. 1. Using as input just a skeletal motion (1st column), we propose an observation-driven neural-network-based solution to add soft-tissue nonlinear dynamics to skinned mesh sequences. Standard linear blend skinning techniques (2nd column) cannot reproduce non-rigid surface dynamics, as shown in the breast area of this jump motion. Our neural network regresses dynamic blendshapes (columns 4,5) that add nonlinear soft-tissue dynamic effects to skinned meshes (3rd column).

We present a novel method to enrich existing vertex-based human body models by adding soft-tissue dynamics. Our model learns to predict per-vertex 3D offsets, referred to as dynamic blendshapes, that reproduce nonlinear mesh deformation effects as a function of pose information. This enables the synthesis of realistic 3D mesh animations, including soft-tissue effects, using just skeletal motion. At the core of our method there is a neural network regressor trained on high-quality 4D scans from which we extract pose, shape and soft-tissue information. Our regressor uses a novel nonlinear subspace, which we build using an autoencoder, to efficiently compact soft-tissue dynamics information. Once trained, our method can be plugged to existing vertex-based skinning methods with little computational overhead (<10ms), enabling real-time nonlinear dynamics. We qualitatively and quantitatively evaluate our method, and show compelling animations with soft-tissue effects, created using publicly available motion capture datasets.

CCS Concepts: • **Computing methodologies** → **Neural networks**; **Animation**; *Motion processing*;

Additional Key Words and Phrases: character animation, neural networks, mesh deformation, skinning, soft-tissue

Authors' addresses: Dan Casas, Universidad Rey Juan Carlos, Madrid, dan.casas@urjc.es; Miguel A. Otaduy, Universidad Rey Juan Carlos, Madrid, miguel.otaduy@urjc.es.

2577-6193/2018/5-ART10 $15.00
https://doi.org/10.1145/3203187

ACM Reference Format:
Dan Casas and Miguel A. Otaduy. 2018. Learning Nonlinear Soft-Tissue Dynamics for Interactive Avatars. *Proc. ACM Comput. Graph. Interact. Tech.* 1, 1, Article 10 (May 2018), 15 pages. https://doi.org/10.1145/3203187

1 INTRODUCTION

Human body modeling is a long standing goal in computer graphics, and a key component for realistic character animation in video games and films. We wish to build a model that generates highly realistic 3D meshes that look and behave as the human body does. Therefore, such a model must be able to represent different body shapes, deform naturally with pose changes, and incorporate nonlinear surface dynamics that mimic the behavior of soft skin. In interactive applications, there is often an additional goal: for simplicity and efficiency, we wish to control the body model using only its skeletal *pose*, with the surface animation a function of this skeletal pose.

Many computer animation methods define the body surface as a kinematic function of skeletal pose. The early linear blend skinning (LBS) method simply blends rigid transformations of skeletal bones, and its limitations have been addressed by defining more complex transformations and/or blending functions [Jacobson and Sorkine 2011; Kavan et al. 2008; Le and Deng 2014; Vaillant et al. 2013; Wang and Phillips 2002]. More sophisticated approaches incorporate actual body surface data into the models [Anguelov et al. 2005; Chen et al. 2013; Loper et al. 2015; Pishchulin et al. 2017; Pons-Moll et al. 2015]. Pose-dependent displacements can be considered the precursor of these approaches [Lewis et al. 2000].

In data-driven body models, a common strategy is to split the model parameters into *shape* and *pose*, enabling explicit parametric control of each space. However, these models ignore the *nonlinear dynamics* of the body's surface, caused by the oscillation of soft-tissue under fast skeletal motion. Dyna [Pons-Moll et al. 2015] is a notable exception. It learns a function that relates the deformation gradient of each surface triangle to skeletal dynamics, composes the deformation gradients due to various sources of motion, and then reconstructs the resulting animated body surface. While effective, this approach does not match the efficient vertex-based animation pipelines suited for interactive applications.

Data-driven dynamics models have been successfully used in other contexts, such as facial muscle models [Sifakis et al. 2005], coarse cloth dynamics [De Aguiar et al. 2010], or particle-based fluids [Ladický et al. 2015]. As opposed to these works, we target a problem with a combination of high-resolution, accuracy and performance requirements.

We propose a novel solution to enrich existing vertex-based human body models, such as LBS or SMPL [Loper et al. 2015]. Our method regresses dynamic blendshapes to add nonlinear soft-tissue dynamics to the traditional piece-wise rigid meshes. To this end, we train a neural network that predicts 3D offsets from joint angles, their velocities and accelerations, together with a recent history of offset values. For higher efficiency, the latter are compressed using a novel autoencoder that reduces the dimensionality of the per-vertex 3D offsets by two orders of magnitude. We demonstrate that our subspace for soft-tissue dynamics overcomes existing methods, based on PCA [Loper et al. 2015; Pons-Moll et al. 2015], and better captures the nonlinear nature of such data.

Overall, our method shows for the first time data-driven nonlinear soft-tissue real-time dynamics in 3D mesh sequences animated just with publicly available skeletal motion data [CMU 2003; Trumble et al. 2017]. Our main contributions are:

- A neural-network-based solution for real-time nonlinear soft-tissue regression to enrich skinned 3D animated sequences.
- A loss function tailored to learn soft-tissue deformations. We compute and leverage per-vertex rigidity to obtain a better behaved minimization problem.

- A novel autoencoder for dimensionality reduction of the 3D vertex displacements that represent nonlinear soft-tissue dynamics in 3D mesh sequences.

2 RELATED WORK

2.1 Skinning

The most basic form of body surface modeling is Linear Blend Skinning (LBS). This technique, limited to a single body shape, attaches an underlying kinematic skeleton to a 3D mesh, and assigns a set of weights to each vertex, which define how the vertex moves with respect to the skeleton. Despite being largely used in video games and films, LBS has two significant limitations: first, articulated areas often suffer from unrealistic deformations such as bulging or the candy-wrap effect; second, the resulting animations are piece-wise rigid and therefore lack surface dynamics. LBS deformation artifacts have been extensively addressed by different solutions, including dual quaternions [Kavan et al. 2008], implicit skinning [Vaillant et al. 2013], optimized centers of rotation [Le and Hodgins 2016], and example-based methods [Kry et al. 2002; Le and Deng 2012, 2014; Lewis et al. 2000; Wang and Phillips 2002], but these solutions ignore the shortcomings of LBS that result from the lack of motion dynamics.

2.2 Data-Driven Body Models

With 3D capturing, scanning and registration methods becoming accessible and accurate [Bogo et al. 2014, 2017; Budd et al. 2013; Huang et al. 2017], many data-driven body models that leverage real data have been proposed. Pioneering data-driven methods proposed to interpolate artist-generated shapes to create new ones [Sloan et al. 2001], or to deform an articulated model to a set of scans in different poses as a means to predict new poses [Allen et al. 2002]. Inspired by these methods, others proposed statistical body models such as SCAPE [Anguelov et al. 2005] and follow-up works [Chen et al. 2013; Hasler et al. 2009; Hirshberg et al. 2012]. These models are learned from static scans and are capable of representing changes due to pose and shape. However, they cannot cope with deformations due to non-rigid surface dynamics. Moreover, they are based on triangle transformations; therefore, they are more expensive to compute than vertex-based models. Recently, Loper et al. [2015] proposed SMPL, a vertex-based method that significantly improves over previous methods. The strength of SMPL is the computation of pose and shape *blendshapes* that generate very compelling articulated 3D meshes by simply adding vertex displacements to a template mesh.

More recently, we have witnessed data-driven models that can cope with human body dynamics. Dyna [Pons-Moll et al. 2015] is arguably one of the best solutions so far. Dyna accurately models shape, pose and soft-tissue dynamics learned from thousands of 4D scans. However, as SCAPE, it is based on triangle deformations, which hinders the implementation of the method in existing vertex-based pipelines such as LBS. DMLP, an extension of SMPL [Loper et al. 2015], also models dynamics. However, the solution relies on a PCA subspace that hinders the learning of nonlinear deformations. In contrast, we show animations with soft-tissue dynamics using skeletal data from publicly available MoCap datasets [CMU 2003; Trumble et al. 2017], we evaluate quantitatively our approach, and we propose a novel autoencoder to build a richer nonlinear subspace that significantly reduces the dimensionality of dynamic blendshapes.

2.3 Physics-Based Body Models

A strong limitation of purely data-driven models is the inherent difficulty to represent deformations far from the training set. Physics-based models overcome this limitation, but are significantly more complex and usually require a volumetric representation of the model. Kadleček et al. [2016] compute a fully physics-based subject-specific anatomical model, including bones, muscle and

soft-tissue. Kim et al. [2017] combine data-driven and physics-based models to create a layered representation that can reproduce soft-tissue effects. Remarkably, they fit the model to captured 4D scans to find subject-specific physical parameters.

Early works also proposed the use of layered representations consisting of a skeleton that drives physics-based soft-tissue deformations [Capell et al. 2002]. Liu et al. [2013] propose a pose-based plasticity model to obtain skinned deformation around joints. Hahn et al. [2012; 2013] enrich standard LBS animations by simulating the deformation of fat and muscles in the nonlinear subspace induced by the model's rig. Other subspaces for deformations have also been explored for characters [Kim and James 2012; Kry et al. 2002], cloth [De Aguiar et al. 2010; Hahn et al. 2014], and for arbitrary geometries based on incremental deformations [Mukherjee et al. 2016]. Xu and Barbič [2016] use secondary Finite Element Method (FEM) dynamics to efficiently add soft-tissue effects to a rigged character.

We also enrich a skinned model with motion-depending deformations, however we do not use a physics-based model. Instead, our soft-tissue deformation is automatically learned with a neural network trained purely from observations.

2.4 Autoencoders for subspaces

Autoencoders are unsupervised neural networks that approximate an identity mapping by coupling an encoding and decoding block to learn a latent subspace representation. Hinton and Salakhutdinov [2006] introduced them and showed their applicability for *nonlinear* dimensionality reduction in greyscale images. Subsequent research showed their applicability in many fields, including image denoising [Vincent et al. 2008] and face reconstruction [Tewari et al. 2017]. In this work, we exploit their *nonlinear* latent space to efficiently encode soft-tissue dynamics. Variational autoencoders [Kingma and Welling 2013] go one step further and enforce a known probability distribution in the latent space, which enables the generation of new data directly from the latent space. This allows, for example, the synthesis of human motion from high-level controls [Habibie et al. 2017], or the generation of images directly from words [Yan et al. 2016] or from silhouettes [Lassner et al. 2017].

3 SOFT-TISSUE REGRESSION

In this section, we describe our learning-based method to augment a skinning-based character animation with realistic nonlinear soft-tissue dynamics. At runtime, our method takes as input a skeletal animation obtained, for example, using motion capture or by editing a rigged character. For each frame of the skeletal animation, our method produces the animation of the character's surface mesh, including nonlinear soft-tissue dynamics effects, following three major steps: surface skinning, compact soft-tissue encoding, and soft-tissue regression. Figure 2 outlines the preprocessing and runtime pipelines of our method. Our choice of skinning model combines a (static) shape representation $\boldsymbol{\beta}$, the skeletal pose $\boldsymbol{\theta}_t$ for the current frame t, and dynamic soft-tissue displacements $\boldsymbol{\Delta}_t$ to produce the deformed surface mesh $\mathbf{M}_t$. This skinning model is described in Sec. 3.1. As we discuss later, a key insight of our method is to represent dynamic soft-tissue displacements in the undeformed pose space.

A naïve design of dynamic soft-tissue regression would suffer from the curse of dimensionality, due to the large size of the soft-tissue displacement vector. To address this challenge, we propose a compact subspace representation of dynamic soft-tissue displacements, obtained using a nonlinear autoencoder, described in Sec. 3.3. For each frame, the autoencoder encodes dynamic soft-tissue displacements $\boldsymbol{\Delta}_t$ into a compact subspace representation $\underline{\boldsymbol{\Delta}}_t$.

Finally, our proposed learning method solves nonlinear soft-tissue dynamics as a nonlinear regression problem. Modeling soft-tissue dynamics boils down to capturing the nonlinear interplay

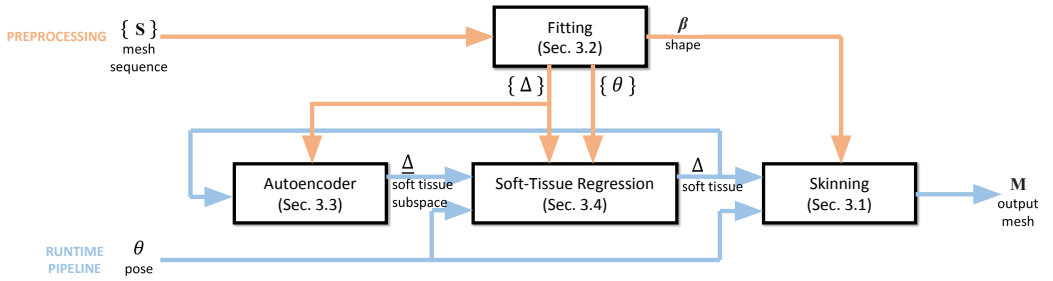


Fig. 2. Overview of the preprocessing and runtime pipelines.

of surface displacements, velocities, and accelerations, with skeletal pose, velocity and acceleration. We model this complex nonlinear function using a neural network, described in Sec. 3.4. The neural network outputs the current dynamic soft-tissue displacement $\boldsymbol{\Delta}_t$, and it takes as input the skeletal pose of the current frame $\boldsymbol{\theta}_t$ and the two previous frames $\boldsymbol{\theta}_{t-1}$ and $\boldsymbol{\theta}_{t-2}$ (to capture skeletal velocity and acceleration), together with the compact soft-tissue displacements of the two previous frames $\underline{\boldsymbol{\Delta}}_{t-1}$ and $\underline{\boldsymbol{\Delta}}_{t-2}$ (to capture soft-tissue velocity and acceleration).

Our method builds on a preprocessing stage, which takes as input a sequence of surface meshes of the character, $\{\mathbf{S}\}$, which span its dynamic behavior. The preprocessing stage involves fitting the surface skinning model and extracting the dynamic soft-tissue deformation, as we describe in Sec. 3.2, together with training the autoencoder and the neural network.

3.1 Skinned Body Model

As a base model we use SMPL [Loper et al. 2015], a data-driven vertex-based linear skinning model that has shown to overcome earlier works such as SCAPE [Anguelov et al. 2005] or BlendSCAPE [Hirshberg et al. 2012] thanks to its simplicity and realism. The key insight in SMPL is the use of *corrective* blendshapes, learned from thousands of 3D body scans, to fix well-known skinning artifacts such as bulging. Formally, SMPL defines a body model surface $\mathbf{M} = M(\boldsymbol{\beta}, \boldsymbol{\theta})$ as

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}) = W(\bar{M}(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \quad (1)$$

$$\bar{M}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \bar{\mathbf{T}} + M_{\mathrm{s}}(\boldsymbol{\beta}) + M_{\mathrm{p}}(\boldsymbol{\theta}), \quad (2)$$

where $W(\bar{\mathbf{T}}, \mathbf{J}, \boldsymbol{\theta}, \mathbf{W})$ is a linear blend skinning function [Magnenat-Thalmann et al. 1988] that computes the *posed* surface vertices of the template $\bar{\mathbf{T}}$ according to the joint locations $\mathbf{J}$, joint angles $\boldsymbol{\theta}$ and blend weights $\mathbf{W}$. The learned functions $M_{\mathrm{s}}(\boldsymbol{\beta})$ and $M_{\mathrm{p}}(\boldsymbol{\theta})$ output vectors of vertex offsets (the *corrective* blendshapes) that, applied to the template $\bar{\mathbf{T}}$, fix classic linear blend skinning artifacts. See Loper et al. [2015] for details.

Our goal is to further deform the vertices of $\bar{\mathbf{T}}$ such that the resulting pose reproduces realistic soft-tissue dynamics. Following SMPL's additive blendshape formulation, we want to find a set of per-vertex 3D offsets $\boldsymbol{\Delta} = \{\boldsymbol{\delta}_i\}_{i=0}^{V-1}$ (which we refer to as *dynamic* blendshape) that added to the template model $\bar{\mathbf{T}}$ will produce the desired deformation to the posed 3D mesh. We therefore extend the body model with an extra blendshape.

$$\bar{M}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \bar{\mathbf{T}} + M_{\mathrm{s}}(\boldsymbol{\beta}) + M_{\mathrm{p}}(\boldsymbol{\theta}) + M_{\mathrm{d}}(\boldsymbol{\gamma}), \quad (3)$$

where $M_{\mathrm{d}}(\boldsymbol{\gamma}) = \boldsymbol{\Delta}$ is a function that regresses the per-vertex offsets $\boldsymbol{\Delta}$, given a history of skeletal motion and surface dynamics in previous frames $\boldsymbol{\gamma}$; see Sec. 3.4 for details. Note that the use of corrective blendshapes for dynamics was also briefly accounted in DMPL (i.e., Dynamic-SMPL) [Loper

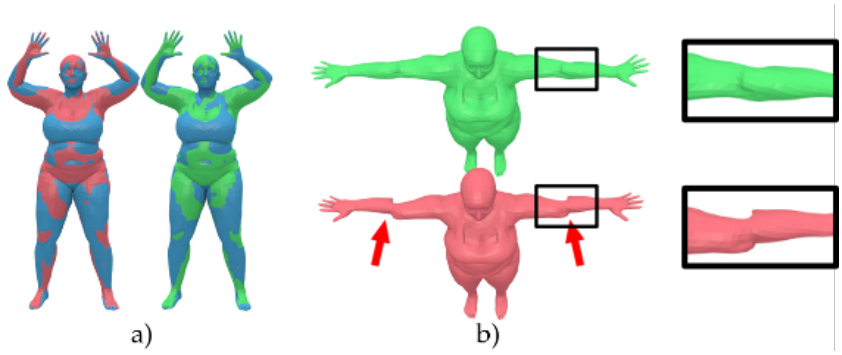


Fig. 3. Fitting parametric model to 4D scan (blue). If minimizing error at the pose state (a, red), unposed scans might suffer from unrealistic deformations (b, red). In contrast, minimizing error at the *unposed* state (a, green [here visualized at pose state]), produces realistic unposes (b, green).

et al. 2015]. However, the DMPL approach relies on a linear PCA subspace, and it was never evaluated nor shown to generalize to arbitrary skeletal motions. In contrast, our solution introduces a novel nonlinear subspace, is easier to train, allows real-time interactions, and it is successfully applied to existing motion-capture datasets.

3.2 Soft-Tissue Ground Truth

As described later in Sec. 3.5, we learn $M_{\mathrm{d}}(\boldsymbol{\gamma})$ from Eq. 3 using a supervised learning method, in particular a neural network. Such machine learning mechanisms, which have recently shown impressive results in many fields including character animation [Fragkiadaki et al. 2015; Holden et al. 2017], are known for their need of large amounts of training data. Ground truth annotated data is usually obtained from observations, manual annotation or physical simulations.

In our case, we leverage recent methods in 4D capture [Bogo et al. 2017; Huang et al. 2017; Pons-Moll et al. 2015; Robertini et al. 2016], which accurately fit and deform a 3D mesh template to reconstruct human performances. Particularly, we use the publicly available dataset of Dyna [Pons-Moll et al. 2015], which captures highly detailed surface deformations at 60fps. Note that here we refer to the aligned 4D scans used in the original paper, *not* the actual Dyna model (which is not public).

Assuming that such 4D scans reproduce the captured surface with negligible error, soft-tissue dynamic deformations can be extracted by fitting the shape-and-pose parametric model defined in Eq. 1 to the scans, and subsequently evaluating the differences between the fitted models and the 4D scans [Kim et al. 2017]. To this end, the parameters $\boldsymbol{\beta}, \boldsymbol{\theta}$ can be found by minimizing

$$\sum_{i=1}^{V} w_i \left\| \mathrm{unpose}(M_i(\boldsymbol{\beta}, \boldsymbol{\theta})) - \mathrm{unpose}(M_i(\mathrm{S}, \boldsymbol{\theta})) \right\|_2^2 , \tag{4}$$

where $\mathrm{unpose}(\cdot)$ is the inverse of the SMPL skinning function (*i.e.*, a function that transforms the mesh to rest pose and removes pose- and shape-corrective blendshapes), $M_i(\cdot)$ is the i^{th} vertex of the mesh, w_i is a weight that is set to high values in rigid parts, and $\mathrm{S} \in \mathbb{R}^{V \times 3}$ is a matrix of vertices of the captured scan. Notice that in contrast to Kim et al. [2017], we perform the minimization on the unposed state. By doing this, we ensure that the ground-truth dynamic blendshapes can be computed directly in the unposed state from optimally matched surfaces. If the minimization is formulated on the posed state, unrealistic deformations are likely to appear when the scan S is unposed if the joint

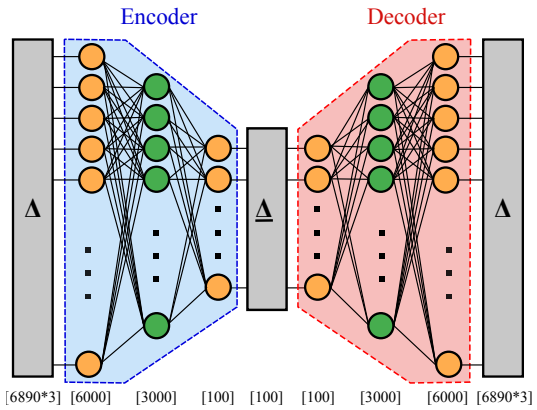


Fig. 4. Autoencoder to compress a dynamic blendshape $\Delta \in \mathbb{R}^{6890 \cdot 3}$ to a compact representation $\underline{\Delta} \in \mathbb{R}^{100}$. Orange and green represent linear and nonlinear activation units, respectively.

positions are not correctly estimated, even if the posed state is accurately matched. The difference is visualized in Fig. 3, where we show the fitted result to a scan $\mathbf{S}$ (blue) minimizing the differences on the posed state (red) and on the unposed state (green). Both results look plausible on the posed state (Fig. 3a), but the unposed scan $\mathbf{S}$ suffers from unrealistic deformations when the minimization is formulated on the posed state (Fig. 3b, red).

We solve Eq. 4 and unpose all frames $\mathbf{S}_t$ of the dataset with per-frame optimized pose $\boldsymbol{\theta}_t$. This step essentially transforms the 4D scans to rest pose, and additionally removes the effect of the SMPL corrective blendshapes due to pose and shape. The residual deformations in the unposed meshes, $\boldsymbol{\Delta}_t \in \mathbb{R}^{V \times 3}$, are the result of soft-tissue dynamic deformation, and constitute our *dynamic* blendshapes. They are formally computed as:

$$\boldsymbol{\Delta}_t = \text{unpose}(M(\boldsymbol{\beta}, \boldsymbol{\theta}_t)) - \text{unpose}(M(\mathbf{S}_t, \boldsymbol{\theta}_t)), \tag{5}$$

Such blendshapes, together with the extracted $\boldsymbol{\theta}_t$,$\boldsymbol{\beta}$, are our ground-truth data, which we use for training the regressor $M_{\text{d}}(\boldsymbol{\gamma})$ from Eq. 3.

3.3 Soft-tissue Autoencoder

Data-driven body models often use an initial dimensionality reduction step to reduce the complexity of data representation. State-of-the-art methods use Principal Component Analysis (PCA) [Anguelov et al. 2005; Feng et al. 2015; Loper et al. 2015; Pons-Moll et al. 2015], a linear method that despite its simplicity successfully reproduces changes due to shape in a significantly lower-dimensional space. Similar linear models have also been used for cloth simulation [De Aguiar et al. 2010], skinning [James and Twigg 2005; Kavan et al. 2010] and physics [Barbič and James 2005].

However, given the highly non-linear nature of the dynamic soft tissue data stored in $\boldsymbol{\Delta}$, a linear model cannot easily represent the deformations in detail. Therefore, in contrast to the earlier works mentioned above, we propose to use an *autoencoder* [Hinton and Salakhutdinov 2006], a nonlinear method that has shown to overcome PCA-based methods in dimensionality reduction capabilities in different fields. Autoencoders approximate an identity mapping by coupling an encoding block with a decoding block to learn a compact intermediate representation, the *latent space*. Particularly, each block consist of a neural network, with different hidden layers and non-linear operators. After training, a forward pass of the encoder converts the input into a compact representation.

Fig. 4 depicts the autoencoder we designed for this work. As input to the encoder, we feed a vectorized version of the dynamic blendshape $\Delta \in \mathbb{R}^{6890 \cdot 3}$. Our encoder consists of three layers with linear, nonlinear, and linear activation functions, respectively, and outputs a vector $\underline{\Delta} \in \mathbb{R}^{100}$. As we demonstrate in Sec 4.1, thanks to the nonlinear activations, we obtain a latent space capable of better reproducing the complexity of soft-tissue dynamics.

3.4 Soft-Tissue Regression

At the core of our method there is a neural network that automatically learns from observations (*i.e.* 4D scans) the function $M_{\mathrm{d}}(\boldsymbol{\gamma}) = \Delta$ in Eq. 3. In particular, $M_{\mathrm{d}}(\boldsymbol{\gamma})$ is parameterized by $\boldsymbol{\gamma} = \{\Delta_{t-1}, \Delta_{t-2}, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t-2}\}$, where $\Delta_{t-1}, \Delta_{t-2}$ are the predicted dynamic blendshapes of previous frames. Notice that $\Delta_t \in \mathbb{R}^{6890 \cdot 3}$ is a prohibitively expensive size for an efficient neural network input, and therefore we reduce the dimensionality using the autoencoder described in Sec. 3.3. This efficiently finds a latent space to encode the nonlinear information. We therefore redefine the input vector to the neural network as $\boldsymbol{\gamma} = \{\underline{\Delta}_{t-1}, \underline{\Delta}_{t-2}, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t-2}\}$.

Notice that another potential choice of regression method could have been a Support Vector Machine (SVM) with a nonlinear kernel. However, this requires carefully picking the kernel function. Additionally, while the size of the neural network is fixed (*i.e.* number of neurons), the size of SVMs usually scales linearly with the number of samples.

3.4.1 Training. Having a set of 4D scans $\mathcal{S} = \{\mathrm{S}_t\}_{t=1}^{T}$, we first extract the dynamic blendshapes Δ_t and the pose and shape parameters $(\boldsymbol{\beta}, \boldsymbol{\theta}_t)$ as described in Sec. 3.2. We then train a single-layer neural network that learns to regress Δ_t from $\boldsymbol{\gamma}$. Each neuron uses a Rectified Linear Unit (ReLU) activation function, a non-linear operator commonly proposed in the machine learning literature due to its simplicity and fast convergence. Note that in order to learn a regressor that *understands* second-order dynamics, we feed a history of the previous dynamic blendshapes to predict the current dynamic blendshape. This makes our predictions much more stable and produces an overall realistic nonlinear behavior, as demonstrated in the supplementary video.

3.4.2 Loss Function. A key part for neural-network training is the choice of an appropriate loss function. In our particular case, we want to minimize the Euclidean distance between vertices of a ground-truth dynamic blendshape $\Delta^{\mathrm{GT}} = \{\boldsymbol{\delta}_i^{\mathrm{GT}}\}_{i=1}^{V}$ and the predicted dynamic blendshape Δ. We therefore minimize the following ℓ_2-norm

$$\mathrm{Loss} = \sum_{i=1}^{V} \left\| w_i^{\mathrm{rig}} \cdot (\boldsymbol{\delta}_i^{\mathrm{GT}} - \boldsymbol{\delta}_i) \right\|_2, \tag{6}$$

where w_i^{rig} is the rigidity weight of the i^{th} vertex, inversely proportional to the vertex stiffness. By adding such weights, we enforce the optimizer to prioritize learning in non-rigid areas, such as breasts and belly, over almost rigid areas, such as the head. We precompute w_i^{rig} automatically from data, also using the input 4D scans, as

$$w_i^{\mathrm{rig}} = \frac{\sum_{t=1}^{T} \left\| \dot{v}_{i,t} - \dot{v}_{i,t-1} \right\|_2}{T}, \tag{7}$$

where $\dot{v}_{i,t}$ is the velocity of the i^{th} vertex of the ground-truth blendshape Δ_t^{GT}, and T the number of frames.

3.5 Networks and Training Details

For each test subject (all obtained from the Dyna dataset), we train both the autoencoder and the soft-tissue regressor on sequences `jumping_jacks`, `light_hopping_loose`, `jiggle_on_toes`,

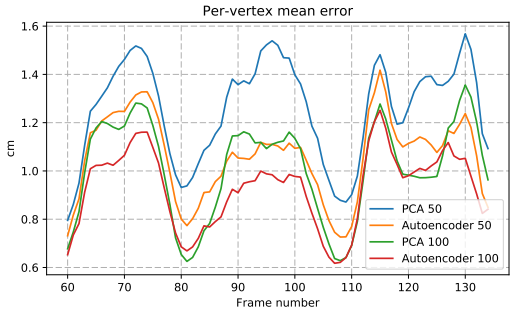


Fig. 5. Quantitative evaluation of the autoencoder on the `50002_running_on_spot` sequence. Our autoencoder for dynamic blendshapes consistently outperforms PCA, allowing the use of smaller subspaces with richer nonlinear reconstruction capabilities.

`one_leg_jump`, and `running_on_spot`, totalling ≈ 1800 frames per subject. As common in machine learning literature, we normalize the input data of each network (*i.e.*, both the autoencoder and the soft-tissue regressor) by subtracting the mean and dividing by the standard deviation.

To train our autoencoder, we use a learning rate of 10^{-5} during 400 epochs. Our soft-tissue regression network, assuming a pose model parameterized by $|\boldsymbol{\theta}| = 75$ DOFs and an autoencoder latent space of 100, consists of a fully-connected single-layer neural network with an input vector $\boldsymbol{\gamma} \in \mathbb{R}^{350}$ $(100 + 100 + 75 + 75 = 350)$ and an output vector $\boldsymbol{\Delta} \in \mathbb{R}^{11\,670}(3890 \cdot 3 = 11\,670)$. We use $\sqrt{|\boldsymbol{\gamma}| \cdot |\boldsymbol{\Delta}|} = 2689$ neurons in the hidden layer, as suggested by Masters [1993].

4 RESULTS AND EVALUATION

We qualitatively and quantitatively evaluate the different stages of our method, including both the autoencoder and the soft-tissue regressor. Additionally, in the supplementary video, we show compelling enriched animations with realistic soft-tissue effects that validate our method.

For training and testing both the autoencoder and the soft-tissue regressor we use the 4D dataset provided in the original Dyna paper [Pons-Moll et al. 2015]. As explained in detail below, all evaluations are carried out using sequences that are not present in the training set.

4.1 Evaluation of the Autoencoder

We evaluate the performance of our autoencoder for dynamic blendshapes by leaving ground-truth sequences `50002_running_on_spot` and `50004_one_leg_jump` out of the training set.

4.1.1 Quantitative Evaluation of the Autoencoder. Fig. 5 plots the per-vertex mean error of the dynamic blendshapes of the sequence `50002_running_on_spot` (*not* used for training), reconstructed with PCA and our autoencoder. Intuitively, higher error in this plot means that the latent space of a particular method fails at reproducing the input mesh. We show results for latent spaces of dimensions 50 and 100 for both PCA and autoencoder. Our autoencoder consistently outperforms PCA when using latent spaces of the same size. Furthermore, notice that our autoencoder with dimension 50 (in orange), performs similarly to PCA with dimension 100 (in green), which demonstrates the richer nonlinear subspace obtained with the proposed autoencoder.

4.1.2 Qualitative Evaluation of the Autoencoder. Fig. 6 shows a reconstructed dynamic blendshape from sequence `50004_one_leg_jump` using PCA and autoencoder, for a range of subspace dimensions. For visualization purposes, we additionally show the reconstruction error using a

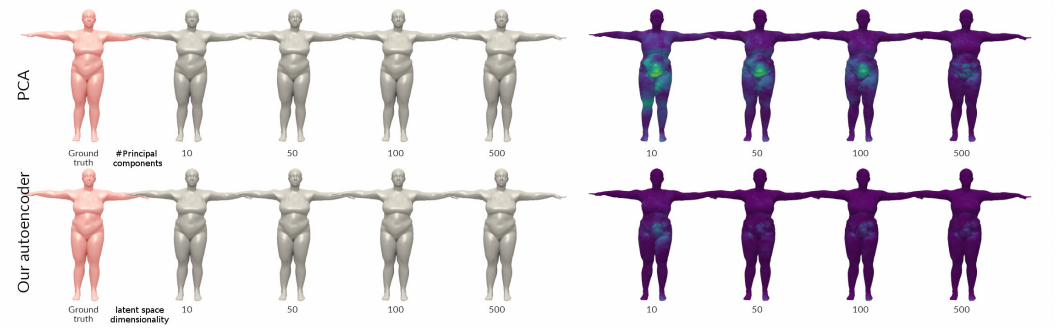


Fig. 6. A dynamic blendshape of a ground-truth scan (pink), and the reconstructions (grey) and colormap errors obtained using different PCA (top) and autoencoder (bottom) subspaces. The frame shown here is part of a jumping motion where the subject just hit the ground, causing a bouncing effect in the belly area. PCA needs a high number of principal components to reproduce such highly dynamic effect; in contrast, our autoencoder successfully reproduces the motion using a significantly smaller latent space. Please see the supplementary video for much clearer visualization.

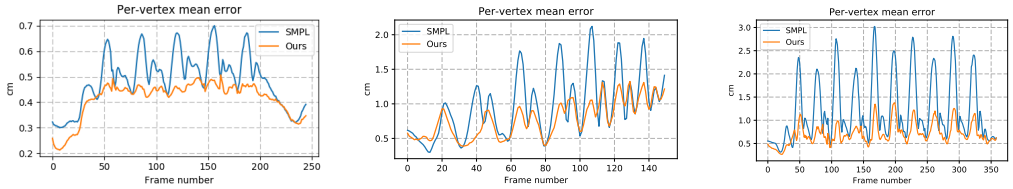
colormap. Our autoencoder consistently outperforms PCA in terms of reconstruction fidelity. Please check the supplementary video for a clearer visualization.

4.2 Evaluation of Soft-Tissue Regression

Our overall goal is to enrich rigged 3D mesh animations created using MoCap data and any skinning technique (LBS, SMPL, etc.). However, results created with such input lack ground-truth data; therefore they do not serve for quantitative evaluation. As an alternaive, we perform quantitative evaluation using a leave-one-out cross-validation strategy on the input 4D scan dataset. We train both the autoencoder and the regressor on all except one sequence of the Dyna dataset [Pons-Moll et al. 2015], and we evaluate the method on the discarded sequence. Notice that 4D scan datasets are rather small with hardly any pose redundancy across sequences (*i.e.* each sequence is a significantly different motion). Therefore, leaving one sequence out of the training set potentially affects the generalization capabilities of the learned model. Despite this observation, we demonstrate robust predictions of soft-tissue dynamics on unseen motions.

In the following, we also provide comparisons to SMPL, the most competitive vertex-based skinning method. Triangle-based skinning methods, such as Dyna, are not directly comparable to ours, as they rely on costly computational methods.

4.2.1 Quantitative Evaluation of Regression. Fig. 7 depicts the mean per-vertex error of our proposed model and SMPL with respect to the ground-truth 4D scans of the Dyna dataset. Note that, for fairness and following our leave-one-out cross-validation strategy, the evaluated sequence in each plot is not part of the training set. In particular, in Fig. 7a we show the mean error over all vertices per frame in the `50004_one_leg_jump` sequence, which results in a mean error of 0.40 ± 0.06 cm, in contrast to the 0.51 ± 0.12 cm error with SMPL. To highlight the improvement in particularly non-rigid areas, such as belly and breasts, in Figs. 7b and 7c we evaluate the mean error only in those areas. Results demonstrate that we improve existing methods by a significant margin. In sequence `50004_running_on_spot`, our method (0.77 ± 0.24cm) significantly outperforms SMPL (1.13 ± 0.52cm); also in sequence `50004_jumping_jacks` (ours 0.71 ± 0.26cm, SMPL 1.22 ± 0.68cm).



(a) 50004_one_leg_jump (b) 50004_running_on_spot (c) 50004_jumping_jacks

Fig. 7. Quantitative evaluation of our model, compared to SMPL. Our results consistently improve over SMPL, reducing the per-vertex mean error with respect to ground-truth scans, specially in frames with highly nonlinear surface motion.

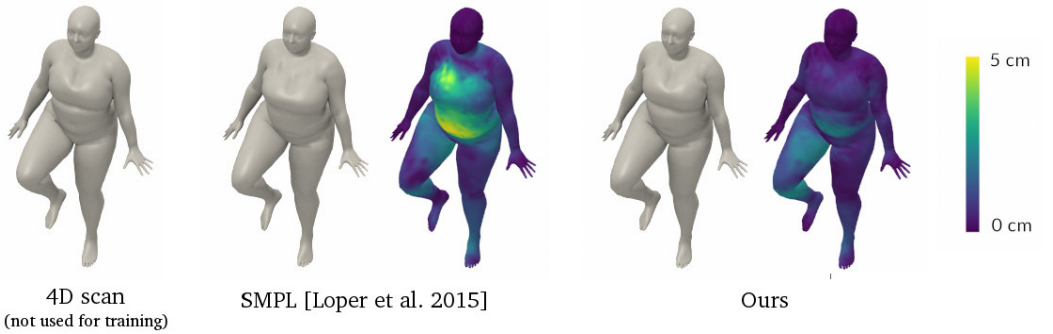


Fig. 8. Qualitative evaluation of our model on the 50004_one_leg_jump sequence. SMPL fails at reproducing nonlinear dynamic motion in the belly and breast area. Our regressed nonlinear dynamics (right) enrich linear models and obtain results closer to ground truth.

4.2.2 Qualitative Evaluation of Regression. We qualitatively evaluate our soft-tissue regression results by visually comparing to ground-truth scans, as well as creating new animations from just skeletal MoCap sequences.

In Fig. 8, as well as in the supplementary video, we provide visual comparisons of our method and SMPL with respect to ground-truth sequences. In particular, Fig. 8 shows one frame of the 50004_one_leg_jump sequence in both plain-geometry and colormap visualization. While SMPL fails at reproducing dynamic details in the belly and breast areas (with errors of up to 5cm), our method successfully reproduces such nonlinear soft-tissue effects. Please see the supplementary video for much clear and in-deep visualizations.

In Figs. 1 and 9, as well as in the supplementary video, we show dynamic sequences created from skeletal MoCap data from publicly available datasets such as CMU [CMU 2003] and Total Capture [Trumble et al. 2017]. Note that we show results for different skeletal hierarchies, which are first converted to the SMPL joint-angle representation. We do this by rigging the Dyna subject template with the input skeletal motion, and then fitting the SMPL model to the rigged animation. Fig. 1 shows how highly non-rigid areas, such as the breasts, are affected by the ongoing motion and are deformed realistically. In the video, we also provide visualizations of the dynamic blendshapes alone, which highlight the degree of soft-tissue deformation. To the best of our knowledge, these are the first examples that augment previously captured skeletal motions with data-driven nonlinear soft-tissue dynamics.

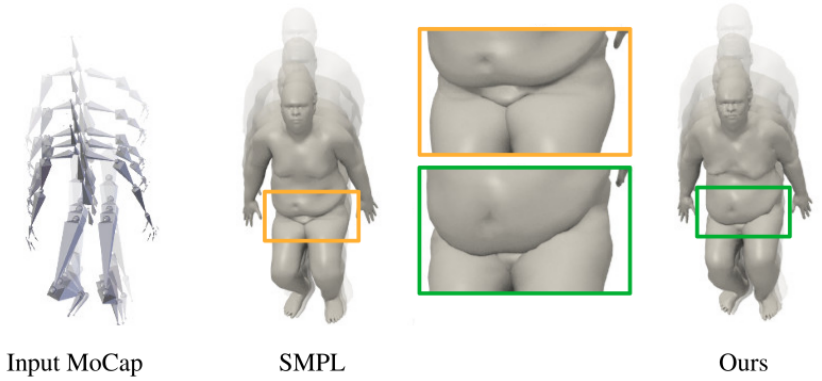


Fig. 9. Mesh sequence created just from MoCap data. SMPL cannot reproduce nonlinear surface dynamics resulting from skeletal motion, such as the belly bounce produced while jumping.

4.3 Run-Time Performance

We have implemented our method in TensorFlow [Abadi et al. 2016] with Adam optimizer [Kingma and Ba 2014], and we have used a regular desktop PC with NVidia GeForce Titan X GPU.

Training of the autoencoder (Sec. 3.3) takes approximately 20min, and training of the soft-tissue regressor (Sec. 3.4) about 40min. Once trained, a forward pass on the encoder takes about 8ms and the soft-tissue regressor about 1ms. Overall, our system performs at real-time rates, including the time budget for standard skinning techniques to produce the input to our method.

5 CONCLUSION

We have presented a novel neural-network-based learning method to enrich rigged characters with nonlinear soft-tissue dynamics. Our body model extends existing vertex-based methods (LBS, SMPL, etc.) by regressing dynamic blendshapes that, added to the model template, enable realistic soft-tissue dynamics. Such dynamic blendshapes are encoded using a new nonlinear subspace that overcomes state-of-the-art (linear) PCA-based subspaces. Both the autoencoder and the soft-tissue regressor are trained solely on captured data. Our model, once trained, is ready to be plugged to any vertex-based skinning method, with little (<10ms) computational overhead, enhancing the realism of the final animation.

As common in learning-based methods, our model struggles when predicting soft-tissue dynamics for poses significantly far from the training set. Future work should study the use of physics-based methods to either extend the training set via simulation, or use physics to further enrich predictions [Kim et al. 2017]. Similarly, our model alone cannot realistically respond to external forces and therefore cannot correctly interact with, for example, physically simulated garments. Future research could study our dynamic body models with soft-tissue from a physical point of view, and subsequently enable physically correct simulation of dressed characters with accurate skin-cloth interaction. This would extend existing solutions in data-driven realistic character animation, which are currently unable to explicitly deal with cloth [Casas et al. 2014].

ACKNOWLEDGMENTS

Dan Casas was supported by a Marie Skłukodowska-Curie Individual Fellowship, grant agreement 707326. This work was supported in part by a grant from the Spanish Ministry of Economy (TIN2015-70799-R).

REFERENCES

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A System for Large-scale Machine Learning. In *Conference on Operating Systems Design and Implementation*. 265–283.

Brett Allen, Brian Curless, and Zoran Popović. 2002. Articulated body deformation from range scan data. In *ACM Transactions on Graphics (TOG)*, Vol. 21. ACM, 612–619.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape Completion and Animation of People. *ACM Transactions on Graphics (TOG)* 24, 3 (2005), 408–416.

Jernej Barbič and Doug L. James. 2005. Real-time subspace integration for St. Venant-Kirchhoff deformable models. *ACM Transactions on Graphics (Proc. of SIGGRAPH)* 24, 3 (2005), 982–990.

Federica Bogo, Javier Romero, Matthew Loper, and Michael J Black. 2014. FAUST: Dataset and evaluation for 3D mesh registration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3794–3801.

Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2017. Dynamic FAUST: Registering Human Bodies in Motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Chris Budd, Peng Huang, Martin Klaudiny, and Adrian Hilton. 2013. Global non-rigid alignment of surface sequences. *International Journal of Computer Vision* 102, 1-3 (2013), 256–270.

Steve Capell, Seth Green, Brian Curless, Tom Duchamp, and Zoran Popović. 2002. Interactive skeleton-driven dynamic deformations. *ACM Transactions on Graphics (Proc. of SIGGRAPH)* 21, 3 (2002), 586–593.

Dan Casas, Marco Volino, John Collomosse, and Adrian Hilton. 2014. 4D Video Textures for Interactive Character Appearance. *Computer Graphics Forum (Proc. Eurographics)* 33, 2 (2014), 371–380.

Yinpeng Chen, Zicheng Liu, and Zhengyou Zhang. 2013. Tensor-based Human Body Modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 105–112.

CMU. 2003. CMU: Carnegie-Mellon Mocap Database. In *http://mocap.cs.cmu.edu*.

Edilson De Aguiar, Leonid Sigal, Adrien Treuille, and Jessica K Hodgins. 2010. Stable Spaces for Real-Time Clothing. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 106.

Andrew Feng, Dan Casas, and Ari Shapiro. 2015. Avatar Reshaping and Automatic Rigging Using a Deformable Model. In *ACM SIGGRAPH Conference on Motion in Games (MIG)*. 57–64.

Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. 2015. Recurrent network models for human dynamics. In *IEEE International Conference on Computer Vision (ICCV)*. 4346–4354.

Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. 2017. Recurrent Variational Autoencoder for Human Motion Synthesis. In *BMVC17*.

Fabian Hahn, Sebastian Martin, Bernhard Thomaszewski, Robert W. Sumner, Stelian Coros, and Markus Gross. 2012. Rig-space physics. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 31, 4 (2012).

Fabian Hahn, Bernhard Thomaszewski, Stelian Coros, Robert W. Sumner, Forrester Cole, Mark Meyer, Tony DeRose, and Markus Gross. 2014. Subspace clothing simulation using adaptive bases. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 105.

Fabian Hahn, Bernhard Thomaszewski, Stelian Coros, Robert W. Sumner, and Markus Gross. 2013. Efficient simulation of secondary motion in rig-space. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 165–171.

Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. 2009. A statistical model of human pose and body shape. In *Computer Graphics Forum (Proc. of Eurographics)*, Vol. 28. 337–346.

Geoffrey E. Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313, 5786 (2006), 504–507.

David A. Hirshberg, Matthew Loper, Eric Rachlin, and Michael J Black. 2012. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *European Conference on Computer Vision (ECCV)*. 242–255.

Daniel Holden, Taku Komura, and Jun Saito. 2017. Phase-functioned Neural Networks for Character Control. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 36, 4 (2017).

Chun-Hao Huang, Benjamin Allain, Edmond Boyer, Jean-Sébastien Franco, Federico Tombari, Nassir Navab, and Slobodan Ilic. 2017. Tracking-by-detection of 3d human shapes: from surfaces to volumes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2017).

Alec Jacobson and Olga Sorkine. 2011. Stretchable and twistable bones for skeletal shape deformation. *ACM Transactions on Graphics (TOG)* 30, 6 (2011).

Doug L. James and Christopher D Twigg. 2005. Skinning mesh animations. *ACM Transactions on Graphics (TOG)* 24, 3 (2005), 399–407.

Petr Kadleček, Alexandru-Eugen Ichim, Tiantian Liu, Jaroslav Křivánek, and Ladislav Kavan. 2016. Reconstructing personalized anatomical models for physics-based body animation. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 35, 6 (2016), 213.

Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. 2008. Geometric skinning with approximate dual quaternion blending. *ACM Transactions on Graphics (TOG)* 27, 4 (2008).

Ladislav Kavan, Peter-Pike Sloan, and Carol O'Sullivan. 2010. Fast and Efficient Skinning of Animated Meshes. *Computer Graphics Forum* 29, 2 (2010), 327–336.

Meekyoung Kim, Gerard Pons-Moll, Sergi Pujades, Sungbae Bang, Jinwwok Kim, Michael Black, and Sung-Hee Lee. 2017. Data-Driven Physics for Human Soft Tissue Animation. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* 36, 4 (2017).

Theodore Kim and Doug L. James. 2012. Physics-based character skinning using multidomain subspace deformations. *IEEE Transactions on Visualization and Computer Graphics* 18, 8 (2012), 1228–1240.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* (2014).

Diederik P. Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114* (2013).

Paul G. Kry, Doug L. James, and Dinesh K. Pai. 2002. Eigenskin: real time large deformation character skinning in hardware. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*. ACM, 153–159.

Lubor Ladický, SoHyeon Jeong, Barbara Solenthaler, Marc Pollefeys, and Markus Gross. 2015. Data-driven Fluid Simulations Using Regression Forests. *ACM Trans. Graph.* 34, 6 (2015), 199:1–199:9.

Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. 2017. A Generative Model of People in Clothing. In *IEEE International Conference on Computer Vision (ICCV)*.

Binh Huy Le and Zhigang Deng. 2012. Smooth Skinning Decomposition with Rigid Bones. *ACM Transactions on Graphics (TOG)* 31, 6 (2012), 199.

Binh Huy Le and Zhigang Deng. 2014. Robust and accurate skeletal rigging from mesh sequences. *ACM Transactions on Graphics (TOG)* 33, 4 (2014).

Binh Huy Le and Jessica K Hodgins. 2016. Real-time skeletal skinning with optimized centers of rotation. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 37.

John P. Lewis, Matt Cordner, and Nickson Fong. 2000. Pose Space Deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Conference on Computer Graphics and Interactive Techniques*. 165–172.

Libin Liu, KangKang Yin, Bin Wang, and Baining Guo. 2013. Simulation and Control of Skeleton-driven Soft Body Characters. *ACM Trans. Graph.* 32, 6, Article 215 (Nov. 2013), 8 pages. https://doi.org/10.1145/2508363.2508427

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (2015), 248:1–248:16.

Nadia Magnenat-Thalmann, Richard Laperrire, and Daniel Thalmann. 1988. Joint-dependent local deformations for hand animation and object grasping. In *Proceedings on Graphics Interface*.

Timothy Masters. 1993. *Practical neural network recipes in C++*. Morgan Kaufmann.

Rajaditya Mukherjee, Xiaofeng Wu, and Huamin Wang. 2016. Incremental Deformation Subspace Reconstruction. *Computer Graphics Forum* 35, 7 (2016), 169–178. https://doi.org/10.1111/cgf.13014

Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. 2017. Building statistical shape spaces for 3D human modeling. *Pattern Recognition* 67 (2017), 276–286.

Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J Black. 2015. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* 34, 4 (2015).

Nadia Robertini, Dan Casas, Helge Rhodin, Hans-Peter Seidel, and Christian Theobalt. 2016. Model-Based Outdoor Performance Capture. In *International Conference on 3D Vision (3DV)*.

Eftychios Sifakis, Igor Neverov, and Ronald Fedkiw. 2005. Automatic Determination of Facial Muscle Activations from Sparse Motion Capture Marker Data. *ACM Trans. Graph.* 24, 3 (July 2005), 417–425. https://doi.org/10.1145/1073204.1073208

Peter-Pike Sloan, Charles F. Rose III, and Michael F. Cohen. 2001. Shape by example. In *Symposium on Interactive 3D Graphics (i3D)*. ACM, 135–143.

Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. 2017. MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*.

Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. 2017. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. In *BMVC17*.

Rodolphe Vaillant, Loïc Barthe, Gaël Guennebaud, Marie-Paule Cani, Damien Rohmer, Brian Wyvill, Olivier Gourmel, and Mathias Paulin. 2013. Implicit skinning: real-time skin deformation with contact modeling. *ACM Transactions on Graphics (TOG)* 32, 4 (2013), 125.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *ACM International Conference on Machine Learning (ICML)*. 1096–1103.

Xiaohuan Corina Wang and Cary Phillips. 2002. Multi-weight enveloping: least-squares approximation techniques for skin animation. In *ACM SIGGRAPH/Eurographics Symposium on Computer animation (SCA)*. 129–138.

Hongyi Xu and Jernej Barbič. 2016. Pose-Space Subspace Dynamics. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 35, 4 (2016).

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2016. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision (ECCV)*. 776–791.