RigNet: Neural Rigging for Articulated Characters

ZHAN XU, YANG ZHOU, and EVANGELOS KALOGERAKIS, University of Massachusetts Amherst
CHRIS LANDRETH and KARAN SINGH, University of Toronto

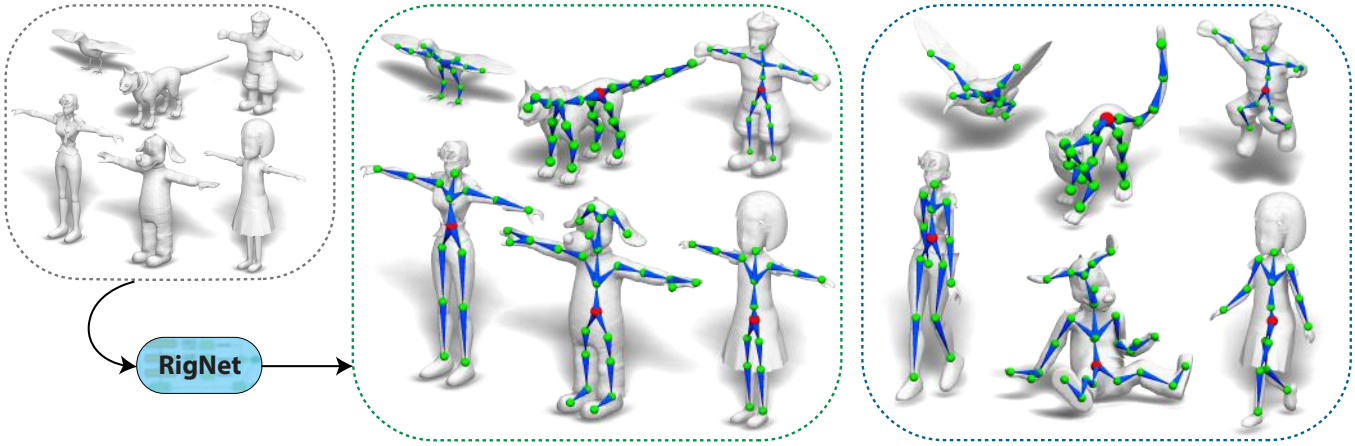

Fig. 1. Given a 3D character mesh, RigNet produces an animation skeleton and skin weights tailored to the articulation structure of the input character. From left to right: input examples of test 3D meshes, predicted skeletons for each of them (joints are shown in green and bones in blue), and resulting skin deformations under different skeletal poses. Please see also our supplementary video: https://youtu.be/J90VETgWIDg

We present *RigNet*, an end-to-end automated method for producing animation rigs from input character models. Given an input 3D model representing an articulated character, *RigNet* predicts a skeleton that matches the animator expectations in joint placement and topology. It also estimates surface skin weights based on the predicted skeleton. Our method is based on a deep architecture that directly operates on the mesh representation without making assumptions on shape class and structure. The architecture is trained on a large and diverse collection of rigged models, including their mesh, skeletons and corresponding skin weights. Our evaluation is three-fold: we show better results than prior art when quantitatively compared to animator rigs; qualitatively we show that our rigs can be expressively posed and animated at multiple levels of detail; and finally, we evaluate the impact of various algorithm choices on our output rigs. [1]

Additional Key Words and Phrases: character rigging, animation skeletons, skinning, neural networks

ACM Reference Format:
Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Landreth, and Karan Singh. 2020. RigNet: Neural Rigging for Articulated Characters. *ACM Trans. Graph.* 39, 4 (to appear), 14 pages. https://doi.org/10.1145/3386569.3392379

[1] Our project page with source code, datasets, and supplementary video is available at https://zhan-xu.github.io/rig-net

1 INTRODUCTION

There is a rapidly growing need for diverse, high-quality, animation-ready characters and avatars in the areas of games, films, mixed Reality and social media. Hand-crafted character "rigs", where users create an animation "skeleton" and bind it to an input mesh (or "skin"), have been the workhorse of articulated figure animation for over three decades. The skeleton represents the articulation structure of the character, and skeletal joint rotations provide an animator with direct hierarchical control of character pose.

We present a deep-learning based solution for automatic rig creation from an input 3D character. Our method predicts both a skeleton and skinning that match animator expectations (Figures 1, 10). In contrast to prior work that fits pre-defined skeletal templates of fixed joint count and topology to input 3D meshes [Baran and Popović 2007], our method outputs skeletons more tailored to the underlying articulation structure of the input. Unlike pose estimation approaches designed for particular shape classes, such as humans or hands [Haque et al. 2016; Huang et al. 2018; Moon et al. 2018; Pavlakos et al. 2017; Shotton et al. 2011; Xu et al. 2017], our approach is not restricted by shape categorization or fixed skeleton structure. Our network represents a generic model of skeleton and skin prediction capable of rigging diverse characters (Figures 1,10).

Predicting an animation skeleton and skinning from an arbitrary single static 3D mesh is an ambitious problem. As shown in Figure 2, animators create skeletons whose number of joints and topology vary drastically across characters depending on their underlying articulation structure. Animators also imbue an implicit understanding of creature anatomy into their skeletons. For example, character spines are often created closer to the back rather than the medial surface or centerline, mimicking human and animal anatomy (Figure 2, cat);

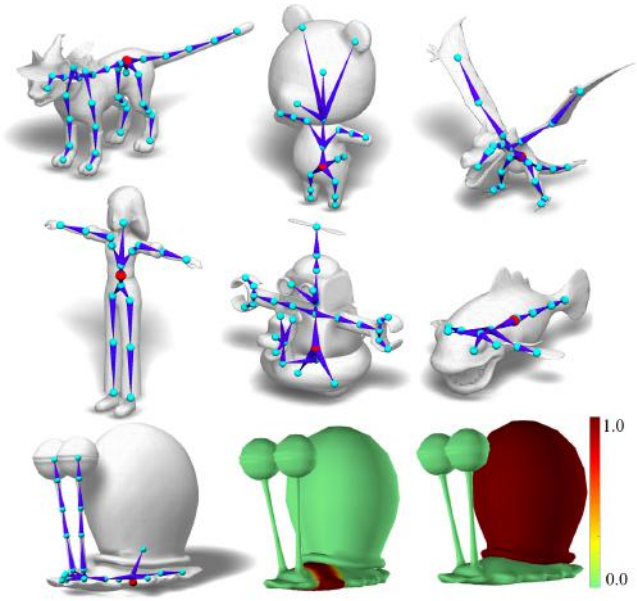

Fig. 2. Examples of skeletons created by animators. In the bottom row, we show a rigged snail, including skinning weights for two of its parts.

they will also likely introduce a proportionate elbow joint into cylindrical arm-like geometry (Figure 2, teddy bear). Similarly when computing skinning weights, animators often perceive structures as highly rigid or smoother (Figure 2, snail). An automatic rigging approach should ideally capture this animators' intuition about underlying moving parts and deformation. A learning approach is well suited for this task, especially if it is capable of learning from a large and diverse set of rigged models.

While animators largely agree on the skeletal topology and layout of joints for an input character, there is also some ambiguity both in terms of number and exact joint placement (Figure 3). For example, depending on animation intent, a hand may be represented using a single wrist joint or at a finer resolution with a hierarchy of hand joints (Figure 3, top row). Spine and tail-like articulations may be captured using a variable number of joints (Figure 3, bottom row). Thus, another challenge for a rigging method is to allow easy and direct control over the level-of-detail for the output skeleton.

To address the above challenges, we designed a deep modular architecture (Figure 4). The first module is a graph neural network, trained to predict an appropriate number of joints and their placement, to capture the articulated mobility of the input character. As skeletal joint resolution can depend on the intended animation task, we provide users an optional parameter that can control the level-of-detail of the output skeleton (Figure 5). A second module learns to predict a hierarchical tree structure (animation skeletons avoid cycles as a design choice) connecting the joints. The output bone structure is a function of joints predicted from the first stage and shape features of the input character. Subsequently, a third module, produces a skinning weight vector per mesh vertex, indicating the degree of influence it receives from different bones. This stage is also based on a graph neural network operating on shape features and intrinsic distances from mesh vertices to the predicted bones.

Our evaluation is three-fold: we show that *RigNet* is better than prior art when quantitatively compared to animator rigs (Tables 1, 2); qualitatively we show our rigs to be expressive and animation-ready (Figure 1 and accompanying video); and technically, we evaluate the impact of various algorithm choices on our output rigs (Tables 3, 4, 5).

In summary, the contribution of this paper is an automated, end-to-end solution to the fundamentally important and challenging problem of character rigging. Our technical contributions include a neural mesh attention and differentiable clustering scheme to localize joints, a graph neural network for learning mesh representations, and a network that learns connectivity of graph nodes (in our case, skeleton joints). Our approach significantly outperforms purely geometric approaches [Baran and Popović 2007], and learning-based approaches that provide partial solutions to our problem i.e., perform only mesh skinning [Liu et al. 2019], or only skeleton prediction for volumetric inputs [Xu et al. 2019].

2 RELATED WORK

In the following paragraphs, we discuss previous approaches for producing animation skeletons, skin deformations of 3D models, and graph neural networks.

Skeletons. Skeletal structures are fundamental representations in graphics and vision [Dickinson et al. 2009; Marr and Nishihara 1978; Tagliasacchi et al. 2016]. Shape skeletons vary in concept from precise geometric constructs like the medial axis representations [Amenta and Bern 1998; Attali and Montanvert 1997; Blum 1973; Siddiqi and Pizer 2008], curvilinear representations or meso-skeletons [Au et al. 2008; Cao et al. 2010; Huang et al. 2013; Singh and Fiume 1998; Tagliasacchi et al. 2009; Yin et al. 2018], to piecewise linear structures [Hilaga et al. 2001; Katz and Tal 2003; Siddiqi et al. 1999; Zhu and Yuille 1996]. Our work is mostly related to animator-centric skeletons [Magnenat-Thalmann et al. 1988], which are designed to capture the mobility of an articulated shape. As discussed in the previous section, apart from shape geometry, the placement of joints and bones in animation skeletons is driven by the animator's understanding of character's anatomy and expected deformations.

The earliest approach to automatic rigging of input 3D models is the pioneering method of "Pinocchio" [Baran and Popović 2007]. Pinocchio follows a combination of discrete and continuous optimization to fit a pre-defined skeleton template to a 3D model, and also performs skinning through heat diffusion. Fitting tends to fail when the input shape structure is incompatible with the selected template. Hand-crafting templates for every possible structural variation of an input character is cumbersome. More recently, inspired by 3D pose estimation approaches [Ge et al. 2018; Haque et al. 2016; Huang et al. 2018; Moon et al. 2018; Newell et al. 2016; Pavlakos et al. 2017; Wan et al. 2018], Xu et al. [Xu et al. 2019] proposed learning a volumetric network for producing skeletons, without skinning, from input 3D characters. Pre-processing the input mesh to a coarser voxel representation can: eliminate surface features (like elbow or knee protrusions) useful for accurate joint detection and placement; alter the input shape topology (like proximal fingers represented as a voxel mitten); or accumulate approximation errors. *RigNet* compares

favorably to these methods (Figure 8, Table 1), without requiring pre-defined skeletal templates, pre-processing or lossy conversion between shape representations.

Skin deformations. A wide range of approaches have also been proposed to model skin deformations, ranging from physics-based methods [Kim et al. 2017; Komaritzan and Botsch 2018, 2019; Mukai and Kuriyama 2016; Si et al. 2015], geometric methods [Bang and Lee 2018; Dionne and de Lasa 2013; Dionne and de Lasa 2014; Jacobson et al. 2011; Kavan et al. 2007; Kavan and Sorkine 2012; Kavan and Žára 2005; Wareham and Lasenby 2008], to data-driven methods that produce skinning from a sequence of examples [James and Twigg 2005; Le and Deng 2014; Loper et al. 2015; Qiao et al. 2018]. Given a single input character, it is common to resort to geometric methods for skin deformation, such as Linear Blend Skinning (LBS) or Dual Quaternion Skinning (DQS) [Kavan et al. 2007; Le and Hodgins 2016] due to their simplicity and computational efficiency. These methods require input skinning weights per vertex which are either interactively painted and edited [Bang and Lee 2018], or automatically estimated based on hand-engineered functions of shape geometry and skeleton [Bang and Lee 2018; Baran and Popović 2007; Dionne and de Lasa 2013; Dionne and de Lasa 2014; Jacobson et al. 2011; Kavan and Sorkine 2012; Wareham and Lasenby 2008]. It is difficult for such geometric approaches to account for any anatomic considerations implicit in input meshes, such as the disparity between animator and geometric spines, or the skin flexibility/rigidity of different articulations.

Data-driven approaches like ours, however, can capture anatomic insights present in animator-created rigs. Neuroskinning [Liu et al. 2019] attempts to learn skinning from an input family of 3D characters. Their network performs graph convolution by learning edge weights within mesh neighborhoods, and outputting vertex features as weighted combinations of neighboring vertex features. Our method instead learns edge feature representations within both mesh and geodesic neighborhoods, and combines them into vertex representations inspired by the edge convolution scheme of [Wang et al. 2019]. Our network input uses intrinsic shape representations capturing geodesic distances between vertices and bones, rather than relying on extrinsic features, such as Euclidean distance. Unlike Neuroskinning, our method does not require any input joint categorization during training or testing. Most importantly, our method proposes a complete solution (skeleton and skinning) with better results (Tables 1, 2).

We note that our method is complementary to physics-based or deep learning methods that produce non-linear deformations, such as muscle bulges, on top of skin deformations [Bailey et al. 2018; Luo et al. 2018; Mukai and Kuriyama 2016], or rely on input bones and skinning weights to compute other deformation approximations [Jeruzalski et al. 2019]. These methods require input bones and skinning weights that are readily provided by our method.

Graph Neural Networks. Graph Neural Networks (GNNs) have become increasingly popular for graph processing tasks [Battaglia et al. 2016; Bruna et al. 2014; Defferrard et al. 2016; Hamilton et al. 2017a,b; Henaff et al. 2015; Kipf and Welling 2016; Li et al. 2016; Scarselli et al. 2009; Wu et al. 2019]. Recently, GNNs have also been proposed for geometric deep learning on point sets [Wang et al. 2019], meshes [Hanocka et al. 2019; Masci et al. 2015], intrinsic or spectral representations [Boscaini et al. 2016; Bronstein et al. 2017; Monti et al. 2017; Yi et al. 2017]. Our graph neural network adapts the operator proposed in [Wang et al. 2019] to perform edge convolutions within mesh-based and geodesic neighborhoods. Our network also weighs and combines representations from mesh topology, local and global shape geometry. Notably, our approach judiciously combines several other neural modules for detecting and connecting joints, with a graph neural network, to provide an integrated deep architecture for end-to-end character rigging.

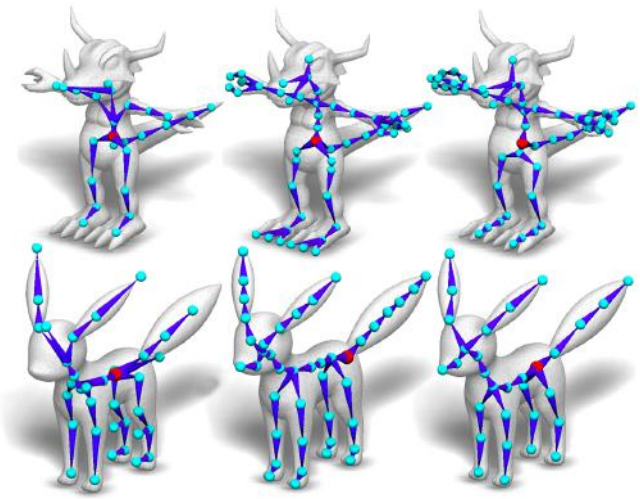

Fig. 3. Models rigged by three different artists. Although they tend to agree on skeleton layout and expected articulation, there is variance in terms of number of joints and overall level-of-detail.

3 OVERVIEW

Given an input 3D mesh of a character, our method predicts an animation skeleton and skinning tailored for its underlying articulation structure and geometry. Both the skeleton and skinning weights are animator-editable primitives that can be further refined through standard modeling and animation pipelines. Our method is based on a deep architecture (Figure 4), which operates directly on the mesh representation. We do not assume known input character class, part structure, or skeletal joint categories during training or testing. Our only assumption is that the input training and test shapes have a consistent upright and frontfacing orientation. Below, we briefly overview the key aspects of our architecture. In Section 4, we explain its stages in more detail.

Skeletal joint prediction. The first module of our architecture is trained to predict the location of joints that will be used to form the animation skeleton. To this end, it learns to displace mesh geometry towards candidate joint locations (Figure 4a). The module is based on a graph neural network, which extracts topology- and geometry-aware features from the mesh to learn these displacements. A key idea of our architecture in this stage is to learn a weight function over the input mesh, a form of neural mesh attention, which is used to reveal which surface areas are more relevant for localizing joints (Figure 4b). Our experiments demonstrate that this leads to more

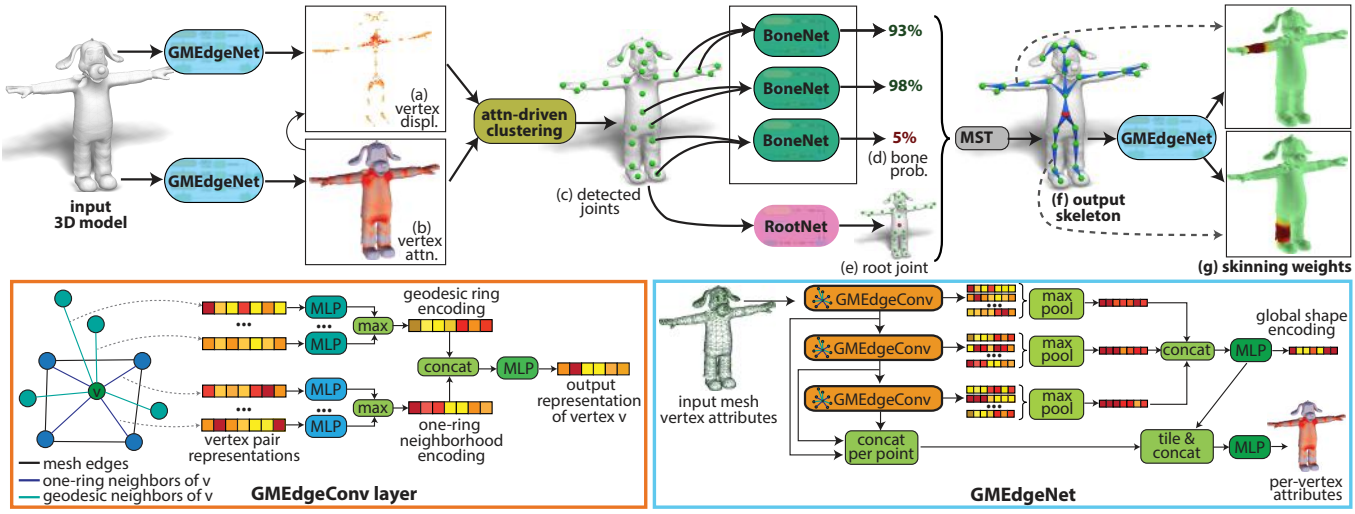

Fig. 4. Top: Pipeline of our method. (a) Given an input 3D model, a graph neural network, namely GMEdgeNet, predicts displacements of vertices towards neighboring joints. (b) Another GMEdgeNet module with separate parameters predicts an attention function over the mesh that indicates areas more relevant for joint prediction (redder values indicate stronger attention - the displaced vertices are also colored according to attention). (c) Driven by the mesh attention, a clustering module detects joints shown as green balls. (d) Given the detected joints, a neural module (BoneNet, see also Figure 6) predicts probabilities for each pair of joints to be connected. (e) Another module (RootNet) extracts the root joint. (f) A Minimum Spanning Tree (MST) algorithm uses the BoneNet and RootNet outputs to form an animation skeleton. (g) Finally, a GMEdgeNet module outputs the skinning weights based on the predicted skeleton. Bottom: Architecture of GMEdgeNet, and its graph convolution layer (GMEdgeConv).

accurate skeletons. The displaced mesh geometry tends to form clusters around candidate joint locations. We introduce a differentiable clustering scheme, which uses the neural mesh attention, to extract the joint locations (Figure 4c).

Since the final animation skeleton may depend on the task or the artists' preferences, our method also allows optional user input in the form of a single parameter to control the level-of-detail, or granularity, of the output skeleton. For example, some applications, like crowd simulation, may not require rigging of small parts (e.g., hands or fingers), while other applications, like FPS games, rigging such parts is more important. By controlling a single parameter through a slider, fewer or more joints are introduced to capture different level-of-detail for the output skeleton (see Figure 5).

Skeleton connectivity prediction. The next module in our architecture learns which pairs of extracted joints should be connected with bones. Our module takes as input the predicted joints from the previous step, including a learned shape and skeleton representation, and outputs a probability representing whether each pair should be connected with a bone or not (Figure 4d). We found that learned joint and shape representations are important to reliably estimate bones, since the skeleton connectivity depends not only on joint locations but also the overall shape and skeleton geometry. The bone probabilities are used as input to a Minimum Spanning Tree algorithm that prioritizes the most likely bones to form a tree-structured skeleton, starting from a root joint picked from another trained neural module (Figure 4e).

Skinning prediction. Given a predicted skeleton (Figure 4f), the last module of our architecture produces a weight vector per mesh vertex indicating the degree of influence it receives from different bones (Figure 4g). Our method is inspired by Neuroskinning [Liu et al. 2019], yet, with important differences in the architecture, bone and shape representations, and the use of volumetric geodesic distances from vertices to bones (as opposed to Euclidean distances).

Training and generalization. Our architecture is trained via a combination of loss functions measuring deviation in joint locations, bone connectivity, and skinning weight differences with respect to the training skeletons. Our architecture is trained on input characters that vary significantly in terms of structure, number and geometry of moving parts e.g., humanoids, bipeds, quadrupeds, fish, toys, fictional characters. Our test set is also similarly diverse. We observe that our method is able to generalize to characters with different number of underlying articulating parts (Figure 10).

4 METHOD

We now explain our architecture (Figure 4) for rigging an input 3D model at test time in detail. In the following subsections, we discuss each stage of our architecture. Then in Section 5, we discuss training.

4.1 Joint prediction

Given an input mesh $\mathcal{M}$, the first stage of our architecture outputs a set of 3D joint locations $\mathbf{t} = \{\mathbf{t}_i\}$, where $t_i \in \mathcal{R}^3$. One particular complication related to this mapping is that the number of articulating parts, and in turn, the number of joints is not the same for all characters. For example, a multiped creature is expected to have more joints than a biped. We use a combination of regression and adaptive clustering to solve for the joint locations and their number. In the regression step, the mesh vertices are displaced towards their nearest candidate joint locations. This step results in accumulating points near joint locations (Figure 4a). The second step localizes the joints by clustering the displaced points and setting the cluster centers

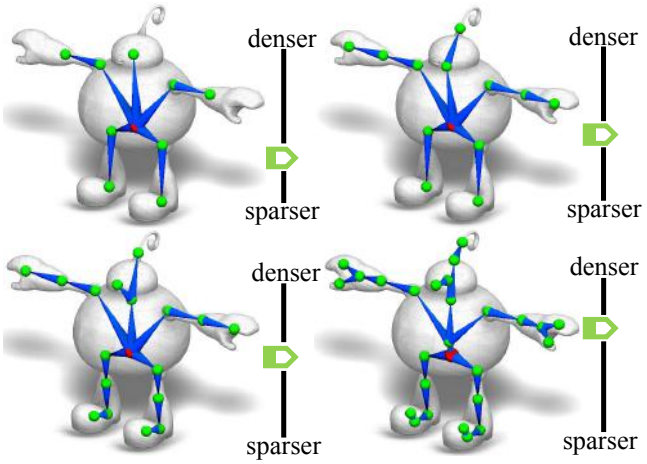

Fig. 5. Effect of increasing the bandwidth parameter that controls the level-of-detail, or granularity, of our predicted skeleton.

as joint locations (Figure 4b). The number of resulting clusters is determined adaptively according to the underlying point density and learned clustering parameters. Performing clustering without first displacing the vertices fails to extract reasonable joints, since the original position of mesh vertices is often far from joint locations. In the next paragraphs, we explain the regression and clustering steps.

Regression. In this step, the mesh vertices are regressed to their nearest candidate joint locations. This is performed through a learned neural network function that takes as input the mesh $\mathcal{M}$ and outputs *vertex displacements*. Specifically, given the original mesh vertex locations $\mathbf{v}$, our displacement module f_d outputs perturbed points $\mathbf{q}$:

$$\mathbf{q} = \mathbf{v} + f_d(\mathcal{M}; \mathbf{w}_d) \tag{1}$$

where $\mathbf{w}_d$ are learned parameters of this module. Figure 4a visualizes displaced points for a characteristic example. This mapping is reminiscent of P2P-Net [Yin et al. 2018] that learns to displace surface points across different domains e.g., surface points to meso-skeletons. In our case, the goal is to map mesh vertices to joint locations. An important aspect of our setting is that not all surface points are equally useful to determine joint locations e.g., the vertices located near the elbow region of an arm are more likely to reveal elbow joints compared to other vertices. Thus, we also designed a neural network function f_a that outputs an attention map which represents a confidence of localizing a joint from each vertex. Specifically, the attention map $\mathbf{a} = \{a_v\}$ includes a scalar value per vertex, where $a_v \in [0, 1]$, and is computed as follows:

$$\mathbf{a} = f_a(\mathcal{M}; \mathbf{w}_a) \tag{2}$$

where $\mathbf{w}_a$ are learned parameters of the attention module. Figure 4b visualizes the map for a characteristic example.

Module internals. Both displacement and attention neural network modules operate on the mesh graph. As we show in our experiments, operating on the mesh graph yields significantly better performance compared to using alternative architectures that operate on point-sampled representations [Yin et al. 2018] or volumetric representations [Xu et al. 2019]. Our networks builds upon the edge convolution proposed in [Wang et al. 2019], also known as 'EdgeConv". Given feature vectors $\mathbf{X} = \{\mathbf{x}_v\}$ at mesh vertices, the output of an EdgeConv operation at a vertex is a new feature vector encoding its local graph neighborhood: $\mathbf{x}'_v = \max_{u \in \mathcal{N}(v)} MLP(\mathbf{x}_v, \mathbf{x}_u - \mathbf{x}_v; \mathbf{w}_{mlp})$ where MLP denotes a learned multi-layer perceptron, $\mathbf{w}_{mlp}$ are its learned parameters, and $\mathcal{N}(v)$ is the graph neighborhood of vertex v. Defining a proper graph neighborhood for our task turned out to be fruitful. One possibility is to simply use one-ring vertex neighborhoods for edge convolution. We instead found that this strategy makes the network sensitive to the input mesh tessellation and results in lower performance. Instead, we found that it is better to define the graph neighborhood of a vertex by considering both its one-ring mesh neighbors, and also the vertices located within a geodesic ball centered at it. We also found that it is better to learn separate MLPs for mesh and geodesic neighborhoods, then concatenate their outputs and process them through another MLP. In this manner, the networks learn to weigh the importance of topology-aware features over more geometry-aware ones. Specifically, our convolution operator, called GMEdgeConv (see also Figure 4, bottom) is defined as follows:

$$\mathbf{x}_{v,m} = \max_{u \in \mathcal{N}_m(v)} MLP(\mathbf{x}_v, \mathbf{x}_u - \mathbf{x}_v; \mathbf{w}_m) \tag{3}$$

$$\mathbf{x}_{v,g} = \max_{u \in \mathcal{N}_g(v)} MLP(\mathbf{x}_v, \mathbf{x}_u - \mathbf{x}_v; \mathbf{w}_g) \tag{4}$$

$$\mathbf{x}'_v = MLP(concat(\mathbf{x}_{v,m},\ \mathbf{x}_{v,g}); \mathbf{w}_c) \tag{5}$$

where $\mathcal{N}_m(v)$ are the one-ring mesh neighborhoods of vertex v, $\mathcal{N}_g(v)$ are the vertices from its geodesic ball. In all our experiments, we used a ball radius $r = 0.06$ of the longest dimension of the model, which is tuned through grid search in a hold-out validation set. The weights $\mathbf{w}_m$, $\mathbf{w}_g$, and $\mathbf{w}_c$ are learned parameters for the above MLPs. We note that we experimented with the attention mechanism proposed in [Liu et al. 2019], yet we did not find any significant improvements. This is potentially due to the fact that EdgeConv already learns edge representations based on the pairwise functions of vertex features, which may implicitly encode edge importance.

Both the vertex displacement and attention modules start with the vertex positions as input features. They share the same internal architecture, which we call GMEdgeNet (see also Figure 4, bottom). GMEdgeNet stacks three GMEdgeConv layers, each followed with a global max-pooling layer. The representations from each pooling layer are concatenated to form a global mesh representation. The output per-vertex representations from all GMEdgeConv layers, as well as the global mesh representation, are further concatenated, then processed through a 3-layer MLP. In this manner, the learned vertex representations incorporate both local and global information. In the case of the vertex displacement module, the feature representation are transformed to 3D displacements per each vertex through another MLP. In the case of the vertex attention module, the per-vertex feature representations are transformed through a MLP and a sigmoid nonlinearity to produce a scalar attention value per vertex. Both modules use their own set of learned parameters for their GMEdgeConv layers and MLPs. More details about their architecture are provided in the appendix.

Clustering. This step takes as input the displaced points $\mathbf{q}$ along with their corresponding attention values $\mathbf{a}$, and outputs joints. As

shown in Figure 4a, points tend to concentrate in areas around candidate joint locations. Areas with higher point density and greater attention are strong indicators of joint presence. We resort to density-based clustering to detect local maxima of point density and use those as joint locations. In particular, we employ a variant of mean-shift clustering, which also uses our learned attention map. A particular advantage of mean-shift clustering is that it does not explicitly require as input the number of target clusters.

In classical mean-shift clustering [Cheng 1995], each data point is equipped with a kernel function. The sum of kernel functions results in a continuous density estimate, and the local maxima (modes) correspond to cluster centers. Mean-shift clustering is performed iteratively; at each iteration, all points are shifted towards density modes. In our implementation, the kernel is also modulated by the vertex attention. In this manner, points with greater attention influence the estimation of density more. Specifically, at each mean-shift iteration, each points is displaced according to the vector:

$$\mathbf{m}_v = \frac{\sum_u a_u \cdot K(\mathbf{q}_u - \mathbf{q}_v, h) \cdot \mathbf{q}_u}{\sum_u a_u \cdot K(\mathbf{q}_u - \mathbf{q}_v, h)} - \mathbf{q}_v \tag{6}$$

where $K(\mathbf{q}_u - \mathbf{q}_v, h) = max(1 - ||\mathbf{q}_u - \mathbf{q}_v||^2/h^2, 0)$ is the Epanechnikov kernel with learned bandwidth h. We found that the Epanechnikov kernel produces better clustering results than a Gaussian kernel or a triangular kernel. The mean-shift iterations are implemented through a recurrent module in our architecture, similarly to the recurrent pixel grouping in Kong and Fowlkes [2018], which also enables training of the bandwidth through backpropagation.

At test time, we perform mean-shift iterations until convergence (i.e., no point is shifted for a Euclidean distance more than 10^{-3}). As a result, the shifted points "collapse" into distinct modes (Figure 4c). To extract these modes, we start with the point with highest density, and remove all its neighbors within radius equal to the bandwidth h. This point represents a mode, and we create a joint at its location. Then we proceed by finding the point with the second largest density among the remaining ones, suppress its neighbors, and create another joint. This process continues until no other points remain. The output of the step are the modes that correspond to the the set of detected joints $\mathbf{t} = \{\mathbf{t}_i\}$.

User control. Since animators may prefer to have more control over the placement of joints, we allow them to override the learned bandwidth value, by interactively manipulating a slider controlling its value (Figure 5). We found that modifying the bandwidth directly affects the level-of-detail of the output skeleton. Lowering the bandwidth results in denser joint placement, while increasing it results in sparser skeletons. We note that the bandwidth cannot be set to arbitrary values e.g., a zero bandwidth value will cause each displaced vertex to become a joint. In our implementation, we empirically set an editable range from 0.01 to 0.1. The resulting joints can be processed by the next modules of our architecture to produce the bone connectivity and skinning based on their updated positions.

Symmetrization. 3D characters are often modeled based on a neutral pose (e.g., "T-pose"), and as a result their body shapes usually have bilateral symmetry. In such cases, we symmetrize joint prediction by reflecting the displaced points $\mathbf{q}$ and attention map $\mathbf{a}$ according to the global bilateral symmetry plane before performing clustering. As a result, the joint prediction is more robust to any small inconsistencies produced in either side.

4.2 Connectivity prediction

Given the joints extracted from the previous stage, the connectivity prediction stage determines how these joints should be connected to form the animation skeleton. At the heart of this stage lies a learned neural module that outputs the probability of connecting each pair of joints via a bone. These pairwise bone probabilities are used as input to Prim's algorithm that creates a Minimum Spanning Tree (MST) representing the animation skeleton. We found that using these bone probabilities to extract the MST resulted in skeletons that agree with animator-created ones more in topology compared to simpler schemes e.g., using Euclidean distances between joints (see Figure 7 and experiments). In the following paragraphs, we explain the module for determining the bone probabilities for each pair of joints, then we discuss the cost function used for creating the MST.

Bone module. The bone module, which we call "BoneNet", takes as input our predicted joints $\mathbf{t}$ along with the input mesh $\mathcal{M}$, and outputs the probability $p_{i,j}$ for connecting each pair of joints via a bone. By processing all pairs of joints through the same module, we extract a pairwise matrix representing all candidate bone probabilities. The architecture of the module is shown in Figure 6. For each pair of joints, the module processes three representations that capture global shape geometry, skeleton geometry, and features from the input pair of joints. In our experiments, we found that this combination offered the best bone prediction performance. More specifically, BoneNet takes as input: (a) a 128-dimensional representation $\mathbf{g}_s$ encoding global shape geometry, which is extracted from the max-pooling layers of GMEdgeNet (see also Figure 4, bottom), (b) a 128-dimensional representation $\mathbf{g}_t$ encoding the overall skeleton geometry by treating joints as a collection of points and using a learned PointNet to produce it [Qi et al. 2017], and (c) a representation encoding the input pair of joints. To produce this last representation, we first concatenate the positions of two joints $\{\mathbf{t}_i, \mathbf{t}_j\}$, their Euclidean distance $d_{i,j}$, and another scalar $o_{i,j}$ capturing the proportion of the candidate bone lying in the exterior of the mesh. The Euclidean distance and proportion are useful indicators of joint connectivity: the smaller the distance between two joints, the more likely is a bone between them. If the candidate bone protrudes significantly outside the shape, then it is less likely to choose it for the final skeleton. We transform the raw features $[\mathbf{t}_i, \mathbf{t}_j, d_{i,j}, o_{i,j}]$ into a 256-dimensional *bone representation* $\mathbf{f}_{i,j}$ through a MLP. The bone probability is computed via a 2-layer MLP operating on the concatenation of these three representations, followed by a sigmoid:

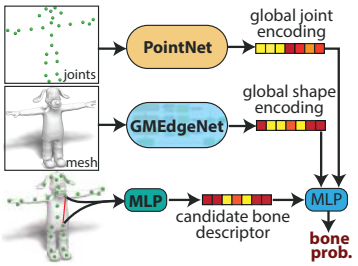

Fig. 6. BoneNet architecture.

$$p_{i,j} = sigmoid\big(MLP(\mathbf{f}_{i,j}, \mathbf{g}_s, \mathbf{g}_t; \mathbf{w}_b)\big) \tag{7}$$

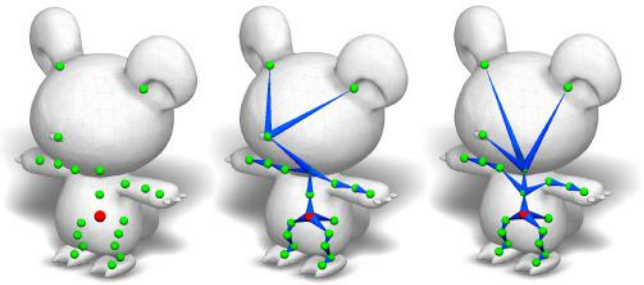

Fig. 7. *Left:* Joints detected by our method. The root joint is shown in red. *Middle:* Skeleton created with Prim's algorithm based on Euclidean distances as edge cost. *Right:* Skeleton created using the negative log of BoneNet probabilities as cost.

where $\mathbf{w}_b$ are learned module parameters. Details about the architecture of BoneNet are provided in the appendix.

Skeleton extraction. The skeleton extraction step aims to infer the most likely tree-structured animation skeleton among all possible candidates. If we consider the choice of selecting an edge in a tree as an independent random variable, the joint probability of a tree is equal to the product of its edge probabilities. Maximizing the joint probability is equivalent to minimizing the negative log probabilities of the edges: $w_{i,j} = -\log p_{i,j}$. Thus, by defining a dense graph whose nodes are the extracted joints, and edges have weights $w_{i,j}$, we can use a MST algorithm to solve this problem. In our implementation, we use Prim's algorithm [Prim 1957]. Any joint can serve as a starting, or root joint for Prim's algorithm. However, since the root joint is used to control the global character's body position and orientation and is important for motion re-targeting tasks, this stage also predicts which joint should be used as root. One common choice is to select the joint closer to the center of gravity for the character. However, we found that this choice is not always consistent with animators' preferences (Figure 2, root nodes in the cat and dragon are further away from their centroids). Instead, we found that the selection of the root joint can also be performed more reliably using a neural module. Specifically, our method incorporates a module, which was call RootNet. Its internal architecture follows BoneNet. It takes as input the global shape representation $\mathbf{g}_s$ and global joint representation $\mathbf{g}_t$ (as in BoneNet). It also takes as input a joint representation $\mathbf{f}_i$ learned through a MLP operating on its location and distance $d_{i,c}$ to the bilateral symmetry plane. The latter feature was driven by the observation that root joints are often placed along this symmetry plane. RootNet outputs the root joint probability as follows:

$$p_{i,r} = softmax\big(MLP(\mathbf{f}_i, \mathbf{g}_s, \mathbf{g}_t; \mathbf{w}_r)\big) \tag{8}$$

where $\mathbf{w}_r$ are learned parameters. At test time, we select the joint with highest probability as root joint to initiate the Prim's algorithm.

4.3 Skinning prediction

After producing the animation skeleton, the final stage of our architecture is the prediction of skinning weights for each mesh vertex to complete the rigging process. To perform skinning, we first extract a mesh representation capturing the spatial relationship of mesh vertices with respect to the skeleton. The representation is inspired by previous skinning methods [Dionne and de Lasa 2013; Jacobson et al. 2011] that compute influences of bones on vertices according to volumetric geodesic distances between them. This mesh representation is processed through a graph neural network that outputs the per-vertex skinning weights. In the next paragraphs, we describe the representation and network.

Skeleton-aware mesh representation. The first step of the skinning stage is to compute a mesh representation $\mathbf{H} = \{\mathbf{h}_v\}$, which stores a feature vector for each mesh vertex v and captures its spatial relationship with respect to the skeleton. Specifically, for each vertex we compute volumetric geodesic distances to all the bones i.e, shortest path lengths from vertex to bones passing through the interior mesh volume. We use a implementation that approximates the volumetric geodesic distances based on [Dionne and de Lasa 2013]; other potentially more accurate approximations could also be used [Crane et al. 2013; Solomon et al. 2014]. Then for each vertex v, we sort the bones according to their volumetric geodesic distance to it, and create an ordered feature sequence $\{\mathbf{b}_{r,v}\}_{r=1...K}$, where r denotes an index to the sorted list of bones. Each feature vector $\mathbf{b}_{r,v}$ concatenates the 3D positions of the starting and end joints of bone r, and the inverse of the volumetric geodesic distance from the vertex v to this bone $(1/D_{r,v})$. The reason for ordering the bones wrt each vertex is to promote consistency in the resulting representation i.e., the first entry represents always the closest bone to the vertex, the second entry represents the second closest bone, and so on. In our implementation, we use the $K = 5$ closest bones selected based on hold-out validation. If a skeleton contains less than K bones, we simply repeat the last bone in the sequence. The final per-vertex representation $\mathbf{h}_v$ is formed by concatenating the vertex position and above ordered sequence $\{\mathbf{b}_{r,v}\}_{r=1...K}$.

Skinning module. The module f_s transforms the above skeleton-aware mesh representation $\mathbf{H}$ to skinning weights $\mathbf{S} = \{\mathbf{s}_v\}$:

$$\mathbf{S} = f_s(\mathbf{H}; \mathbf{w}_s) \tag{9}$$

where $\mathbf{w}_s$ are learned parameters. The skinning network follows GMEdgeNet. The last layer outputs a 1280-dimensional per-vertex feature vector, which is transformed to a per-vertex skinning weight vector $\mathbf{s}_v$ through a learned MLP and a softmax function. This ensures that the skinning weights for each vertex are positive and sum to 1. The entries of the output skinning weight vector $\mathbf{s}_v$ are ordered according to the volumetric geodesic distance of the vertex v to the corresponding bones.

5 TRAINING

The goal of our training procedure is to learn the parameters of the networks used in each of the three stages of *RigNet*. Training is performed on a dataset of rigged characters described in Section 6.

5.1 Joint prediction stage training

Given a set of training characters, each with skeletal joints $\hat{\mathbf{t}} = \{\hat{\mathbf{t}}_k\}$, we learn the parameters $\mathbf{w}_a$, $\mathbf{w}_d$, and bandwidth h of this stage such that the estimated skeletal joints approach as closely as possible to the training ones. Since the estimated skeletal joints originate from mesh vertices that collapse into modes after mean shift clustering, we can alternatively formulate the above learning goal as a problem of minimizing the distance of collapsed vertices to nearest training joints

and vice versa. Specifically, we minimize the symmetric Chamfer distance between collapsed vertices $\{\mathbf{t}_v\}$ and training joints $\{\hat{\mathbf{t}}_k\}$:

$$L_{cd}(\mathbf{w}_a, \mathbf{w}_d, h) = \frac{1}{V}\sum_{v=1}^{V}\min_k ||\mathbf{t}_v - \hat{\mathbf{t}}_k|| + \frac{1}{K}\sum_{k=1}^{K}\min_v ||\mathbf{t}_v - \hat{\mathbf{t}}_k|| \quad (10)$$

The loss is summed over the training characters (we omit this summation for clarity). We note that this loss is differentiable wrt all the parameters of the joint prediction stage, including the bandwidth. The mean shift iterations of Eq. 6 are differentiable with respect to the attention weights and displaced points. This allows us to back-propagate joint location error signal to both the vertex displacement and attention network. The Epanechnikov kernel in mean-shift is also a quadratic function wrt the bandwidth, which makes it possible to learn the bandwidth efficiently through gradient descent. Learning converged to a value of $h = 0.057$ based on our training dataset.

We also found that adding supervisory signal to the vertex displacements before clustering helped improving training speed and joint detection performance (see also experiments). To this end, we minimize Chamfer distance between displaced points and ground-truth joints, favoring tighter clusters:

$$L'_{cd}(\mathbf{w}_d) = \frac{1}{V}\sum_{v}\min_k ||\mathbf{q}_v - \hat{\mathbf{t}}_k|| + \frac{1}{K}\sum_{k}\min_v ||\mathbf{q}_v - \hat{\mathbf{t}}_k|| \quad (11)$$

This loss affects only the parameters $\mathbf{w}_d$ of the displacement module. Finally, we found that adding supervision to the vertex attention weights also offered a performance boost, as discussed in our experiments. This loss is driven by the observation that the displacement of vertices located closer to joints are more helpful to localize them more accurately. Thus, for each training mesh, we find vertices closest to each joint at different directions perpendicular to the bones. Then we create a binary mask $\hat{\mathbf{m}}$ whose values are equal to 1 for these closest vertices, and 0 for the rest. We use cross-entropy to measure consistency between these masks and neural attention:

$$L_m(\mathbf{w}_a) = \hat{\mathbf{m}}\log\mathbf{a} + (1-\hat{\mathbf{m}})\log(1-\mathbf{a})$$

Edge dropout. During training of GMEdgeNet, for each batch, we randomly select a subset of edges within geodesic neighborhoods (in our implementation, we randomly select subsets up to 15 edges). This sampling strategy can be considered as a form of mesh edge dropout. We found that it improved performance since it simulates varying vertex sampling on the mesh, making the graph network more robust to different tessellations.

Training implementation details. We first pre-train the parameters $\mathbf{w}_a$ of attention module with the loss L_m alone. We found that bootstrapping the attention module with this pre-training helped with the performance (see also experiments). Then we fine-tune $\mathbf{w}_a$, and train the parameters $\mathbf{w}_d$ of the displacement module and the bandwidth h using the combined loss: $L_{cd}(\mathbf{w}_a, \mathbf{w}_d, h) + L'_{cd}(\mathbf{w}_d)$. For fine-tuning, we use the Adam optimizer with a batch size of 2 training characters, and learning rate 10^{-6}.

5.2 Connectivity stage training

Given a training character, we form the adjacency matrix encoding the connectivity of the skeleton i.e., $\hat{p}_{ij} = 1$ if two training joints i and j are connected, and $\hat{p}_{ij} = 0$ otherwise . The parameters $\mathbf{w}_b$ of the BoneNet are learned using binary cross-entropy between the training adjacency matrix entries and the predicted probabilities $p_{i,j}$:

$$L_m(\mathbf{w}_a) = \sum_{i,j}\hat{p}_{ij}\log p_{i,j} + (1-\hat{p}_{ij})\log(1-p_{i,j})$$

The BoneNet parameters are learned using the probabilities $p_{i,j}$ estimated for training joints rather than the predicted ones of the previous stage. The reason is that the training adjacency matrix is defined on training joints (and not on the predicted ones). We tried to find correspondences between the predicted joints and the training ones using the Hungarian method, then transfer the training adjacencies to pairs of matched joints. However, we did not observe significant improvements by doing this potentially due to matching errors. Finally, to train the parameters $\mathbf{w}_r$ of the network used to extract the root joint, we use the softmax loss for classification.

Training implementation details. Training BoneNet has an additional challenge due to class imbalance problem: out of all pairs of joints, only few are connected. To deal with this issue, we adopt the online hard-example mining approach from [Shrivastava et al. 2016]. For both networks, we employ the Adam optimizer with batch size 12 and learning rate 10^{-3}.

5.3 Skinning stage training

Given a set of training characters, each with skin weights $\hat{\mathbf{S}} = \{\hat{\mathbf{s}}_v\}$, we train the parameters $\mathbf{w}_s$ of our skinning network so that the estimated skinning weights $\mathbf{S} = \{\mathbf{s}_v\}$ agree as much as possible with the training ones. By treating the per-vertex skinning weights as probability distributions, we use cross-entropy as loss to quantify the disagreement between training and predicted distributions for each vertex:

$$L_s(\mathbf{w}_s) = \frac{1}{V}\sum_{v}\sum_{r}\hat{s}_{v,r}\log s_{v,r}$$

As in the case of the connectivity stage, we train the skinning network based on the training skeleton rather than the predicted one, since we do not have skinning weights for it. We tried to transfer skinning weights from the training bones to the predicted ones by establishing correspondences as before, but this did not result in significant improvements.

Training implementation details. To train the skinning network, we use the Adam optimizer with a batch size of 2 training characters, and learning rate 10^{-4}. We also apply the edge dropout scheme during the training of this stage, as in the joint prediction stage.

6 RESULTS

We evaluated our method and alternatives for animation skeleton and skinning prediction both quantitatively and qualitatively. Below we discuss the dataset used for evaluation, the performance measures, comparisons, and ablation study.

Dataset. To train and test our method and alternatives, we chose the "ModelsResource-RigNetv1" dataset of 3D articulated characters from [Xu et al. 2019], which provides a non-overlapping training and test split, and contains diverse characters [2]. Specifically, the

[2]please see also our project page: https://zhan-xu.github.io/rig-net

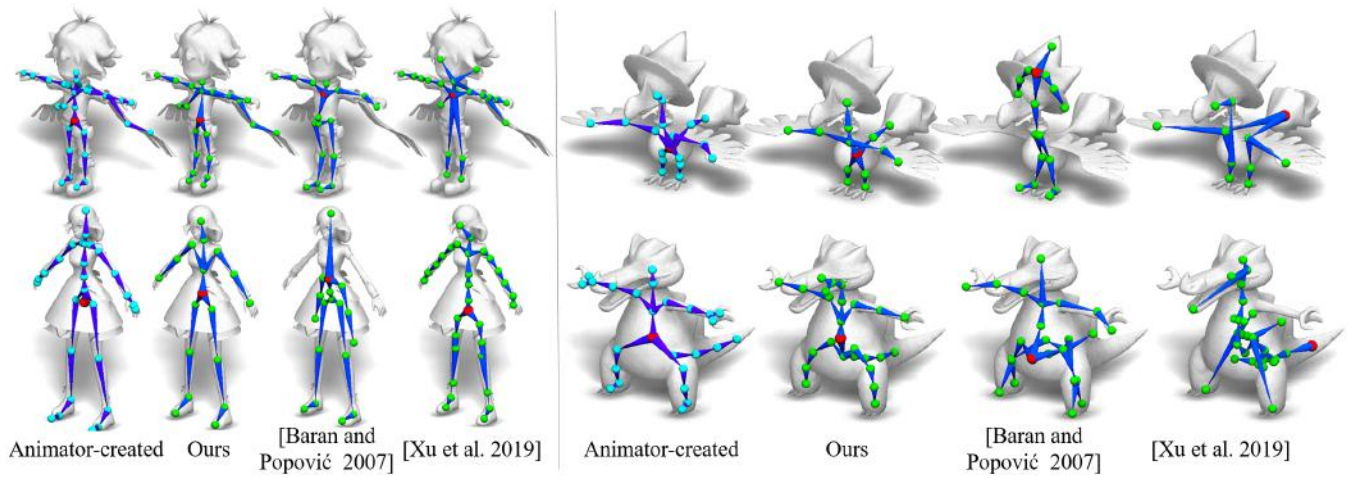

Fig. 8. Comparisons with previous methods for skeleton extraction. For each character, the reference skeleton is shown on the left ("animator-created"). Our predictions tend to agree more with the reference skeletons.

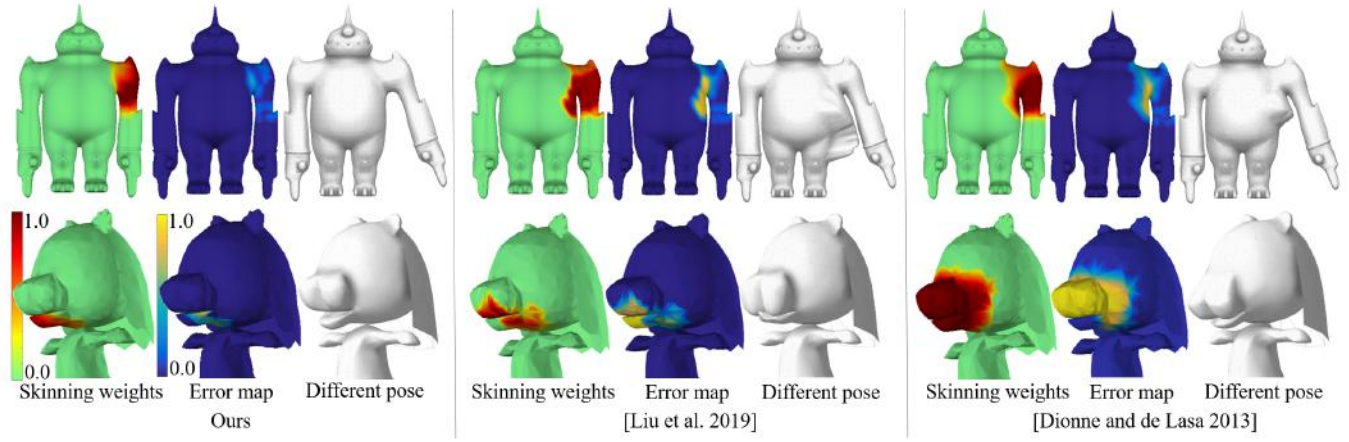

Fig. 9. Comparisons with prior methods for skinning. We visualize skinning weights, L1 error maps, and a different pose (moving right arm for the robot above, and lowering the jaw of the character below). Our method produces lower errors in skinning weight predictions on average.

dataset contains 2703 rigged characters mined from an online repository [Models-Resource 2019], spanning several categories, including humanoids, quadrupeds, birds, fish, robots, toys, and other fictional characters. Each character includes one rig (we note that the multiple rig examples of the two models of Figure 3 were made separately and do not belong to this dataset). The dataset does not contain duplicates, or re-meshed versions of the same character. Such duplicates were eliminated from the dataset. Specifically, all models were voxelized in a binary 88^3 grid, then for each model in the dataset, we computed the Intersection over Union (IoU) with all other models based on their volumetric representation. We eliminated duplicates or near-duplicates whose IoU of volumes was more than 95%. We also manually verified that such re-meshed versions were filtered out. Under the guidance of an artist, we also verified that all characters have plausible skinning weights and deformations. We use a training, hold-out validation, and test split, following a 80%-10%-10% proportion respectively, resulting in 2163 training, 270 hold-out validation, and 270 test characters. Figure 2 shows examples from the training split. The models are consistently oriented and scaled. Meshes with fewer than $1K$ vertices were subdivided; as a result all training and test meshes contained between 1K and 5K vertices. The number of joints per character varied from 3 to 48, and the average is 25.0. The quantitative and qualitative evaluation was performed on the test split of the dataset.

Quantitative evaluation measures. Our quantitative evaluation aims to measure the similarity of the predicted animation skeletons and skinning to the ones created by modelers in the test set (denoted as "reference skeletons" and "reference skinning" in the following paragraphs). For evaluating skeleton similarity, we employ various measures following [Xu et al. 2019]:

(a) *CD-J2J* is the symmetric Chamfer distance between joints. Given a test shape, we measure the Euclidean distance from each predicted joint to the nearest joint in its reference skeleton, then divide with the number of predicted joints. We also compute the Chamfer distance the other way around from the reference skeletal joints to the nearest predicted ones. We denote the average of the two as *CD-J2J*.

(b) *CD-J2B* is the Chamfer distance between joints and bones. The

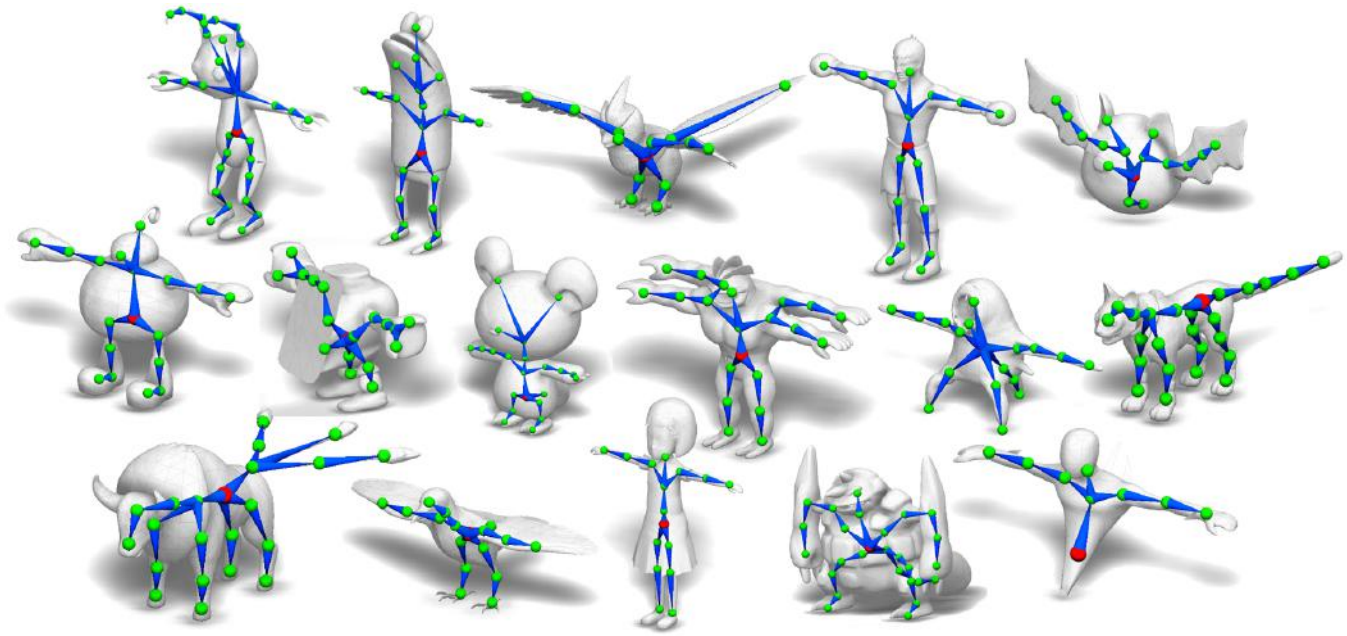

Fig. 10. Predicted skeletons for test models with varying structure and morphology. Our method is able to produce reasonable skeletons even for models that have different number or types of parts than the ones used for training e.g., quadrupeds with three tails.

difference from the previous measure is that for each predicted joint, we compute its distance to the nearest bone point on the reference skeleton. We symmetrize this measure by also computing the distance from reference joints to predicted bones. A low value of *CD-J2B* and a high value of *CD-J2J* mean that the predicted and reference skeletons tend to overlap, yet the joints are misplaced along the bone direction.

(c) *CD-B2B* is the Chamfer distance between bones (line segments). As above, we define it symmetrically. CD-B2B measures similarity of skeletons in terms of bone placement (rather than joints). Ideally, all *CD-J2J*, *CD-J2B*, and *CD-B2B* measures should be low.

(d) *IoU* (Intersection over Union) can also be used to characterize skeleton similarity. First, we find a maximal matching between the predicted and reference joints by using the Hungarian algorithm. Then we measure the number of predicted and reference joints that are matched and whose Euclidean distance is lower then a prescribed tolerance. This is then divided with the total number of predicted and reference joints. By varying the tolerance, we can obtain plots demonstrating IoU for various tolerance levels (see Figure 11). To provide a single, informative value, we set the tolerance to half of the local shape diameter [Shapira et al. 2008] evaluated at each corresponding reference joint. This is evaluated by casting rays perpendicular to the bones connected at the reference joint, finding ray-surface intersections, and computing the joint-surface distance averaged over all rays. The reason for this normalization is that thinner parts e.g, arms have lower shape diameter; as a result, small joint deviations can cause more noticeable misplacement compared to thicker parts like torso.

(e) *Precision & Recall* can also be used here. Precision is the fraction of predicted joints that were matched and whose distance to their nearest reference one is lower than the tolerance defined above. Recall is the fraction of reference joints that were matched and whose distance to their nearest predicted joints is lower than the tolerance. Note that since the number of reference or predicted joints may not be the same. Unmatched predicted joints contribute no precision, and similarly unmatched reference joints contribute no recall.

(f) *TreeEditDist (ED)* is the tree edit distance measuring the topological difference of the predicted skeleton to the reference one. The measure is defined as the minimum number of joint deletions, insertions, and replacements that are necessary to transform the predicted skeleton into the reference one.

	IoU	Prec.	Rec.	CD-J2J	CD-J2B	CD-B2B
Pinocchio	36.5%	38.7%	35.9%	7.2%	5.5%	4.7%
Xu et al. 2019	53.7%	53.9%	55.2%	4.5%	2.9%	2.6%
Ours	**61.6%**	**67.6%**	**58.9%**	**3.9%**	**2.4%**	**2.2%**

Table 1. Comparisons with other skeleton prediction methods.

To evaluate skinning, we use the reference skeletons for all methods, and measure similarity between predicted and reference skinning maps:

(a) *Precision & Recall* are measured by finding the set of bones that influence each vertex significantly, where influence corresponds to a skinning weight larger than a threshold ($1e^{-4}$, as described in [Liu et al. 2019]). Precision is the fraction of influential bones based on the predicted skinning among the ones defined based on the reference skinning. Recall is the fraction of the influential bones based on the reference skinning matching the ones found from the predicted skinning.

(b) *L1-norm* measures the L1 norm of the difference between the predicted skinning weight vector and the reference one for each mesh vertex. We compute the average L1-norm over each test mesh.

(c) *dist* measures the Euclidean distance between the position of vertices deformed based on the reference skinning and the predicted

	Prec.	Rec.	avg L1	avg dist	max dist
BBW	68.3%	77.6 %	0.69	0.0061	0.055
GeoVoxel	72.8%	75.1 %	0.65	0.0057	0.049
NeuroSkinning	76.3%	74.7 %	0.57	0.0053	0.043
Ours	**82.3%**	**80.8%**	**0.39**	**0.0041**	**0.032**

Table 2. Comparisons with other skinning prediction methods.

one. To this end, given a test shape, we generate 10 different random poses, and compute the average and max distance error over the mesh vertices.

All the above skeleton and skinning evaluation measures are computed for each test shape, then averaged over the the test split.

Competing methods. For skeleton prediction, we compare our method with *Pinocchio* [Baran and Popović 2007] and [Xu et al. 2019]. Pinocchio fits a template skeleton for each model. The template is automatically selected among a set of predefined ones (humanoid, short quadruped, tall quadruped, and centaur) by evaluating the fitting cost for each of them, and choosing the one with the least cost. [Xu et al. 2019] is a learning method trained on the same split as ours, with hyper-parameters tuned in the same validation split. For skinning weights prediction, we compare with the Bounded-Biharmonic Weights (BBW) method [Jacobson et al. 2011], NeuroSkinning [Liu et al. 2019] and the geometric method from [Dionne and de Lasa 2013], called "GeoVoxel". For the BBW method, we adopt the implementation from libigl [Jacobson et al. 2018], where the mesh is first tetrahedralized, then the bounded biharmonic weights are computed based on this volume discretization. For NeuroSkinning, we trained the network on the same split as ours and optimized its hyperparameters in the same hold-out validation split. For GeoVoxel, we adopt Maya's implementation [Autodesk 2019] which outputs skinning weights based on a hand-engineered function of volumetric geodesic distances. We set the max influencing bone number, weight pruning threshold, and drop-off parameter through holdout validation in our validation split (3 bones, 0.3 pruning threshold, and 0.5 dropoff).

Comparisons. Table 1 reports the evaluation measures for skeleton extraction between competing techniques. Our method outperforms the rest according to all measures. This is also shown in Fig.11, showing IoU on the y-axis for different tolerance levels (multipliers of local shape diameter) on the x-axis.

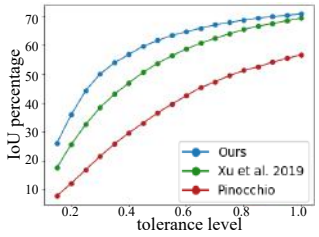

Fig. 11. IoU vs different tolerances.

Figure 8 visualizes reference skeletons and predicted ones for different methods for some characteristic test shapes. We observe that our method tends to output skeletons whose joints and bones are closer to the reference ones. [Baran and Popović 2007] often produces implausible skeletons when the input model has parts (e.g., tail, clothing) that do not correspond well to the used template. [Xu et al. 2019] tends to misplace joints around areas, such as elbows and knees, since voxel grids tend to lose surface detail.

	IoU	Prec.	Rec.	CD-J2J	CD-J2B	CD-B2B
P2PNet-based	40.6%	41.6%	42.0%	6.3%	4.6%	3.8%
No attn	52.4%	50.9%	50.7%	4.6%	3.1%	2.7%
One-ring	59.7%	65.6%	57.4%	4.1%	2.5%	2.4%
No vertex loss	59.3%	58.2%	57.6%	4.2%	2.7%	2.5%
No attn pretrain	60.6%	64.0%	58.1%	4.2%	2.6%	2.4%
Full	**61.6%**	**67.6%**	**58.9%**	**3.9%**	**2.4%**	**2.2%**

Table 3. Joint prediction ablation study

	Class. Acc.	CD-B2B	ED
Euclidean edge cost	61.2%	0.30%	5.0
bone descriptor only	71.9%	0.22%	4.2
bone descriptor+skel. geometry	80.7%	0.12%	2.9
Full stage	**83.7%**	**0.10%**	**2.4**

Table 4. Connectivity prediction ablation study

	Prec	Rec.	avg-L1	avg-dist.	max-dist.
No geod. dist.	80.0%	79.3%	0.41	0.0044	0.054
Ours	**82.3%**	**80.8%**	**0.39**	**0.0041**	**0.032**

Table 5. Skinning prediction ablation study

Table 2 reports the evaluation measures for skinning. Our numerical results are significantly better than BBW, NeuroSkinning, and GeoVoxel according to all the measures. Figure 9 visualizes the skinning weights produced by our method, GeoVoxel, and NeuroSkining that were found to be the best alternatives according to our numerical evaluation. Ours tends to agree more with the artist-specified skinning. On the top example, arms are close to torso in terms of Euclidean distance, and to some degree also in geodesic sense. Both NeuroSkining and GeoVoxel over-extend the skinning weights to a larger area than the arm. In order to match the GeoVoxel's output to the artist-created one, all its parameters need to be manually tuned per test shape, which is laborious. Our method combines bone representations and vertex-skeleton intrinsic distances in our mesh network to produce skinning that better separates articulating parts. In the bottom example, a jaw joint is placed close to the lower lip to control the jaw animation. Most vertices on the front face are close to this joint in terms of both geodesic and Euclidean distances. This results in higher errors for both NeuroSkinning and GeoVoxel, even if the latter is manually tuned. Our method produces a sharper map capturing the part of the jaw.

Ablation study. We present the following ablation studies to demonstrate the influence from different design choices of our method.
(a) *Joint prediction ablation study*: Table 3 presents evaluation of variants of our joint detection stage trained in the same split and tuned in the same hold-out validation split as our original method. We examined the following variants: *"P2PNet-based"* uses the same architecture as P2PNet [Yin et al. 2018], which relies on PointNet [Qi et al. 2017] for displacing points (vertices in our case). After displacement, mean-shift clustering is used to extract joints as in our method. We experimented with the loss from their approach, and also the same loss as in our joint detection stage (excluding the attention mask loss, since P2PNet does not use attention). The latter choice worked better. The architecture was trained and tuned in the same split as ours. *"No attn"* is our method without the attention module,

thus all vertices have the same weight during clustering. *"One-ring"* is our method where GMEdgeConv uses only one-ring neighbors of each vertex without considering geodesic neighborhoods. *"No vertex loss"* does not use vertex displacement supervision with the Chamfer distance loss of Eq. 11 during training. It uses supervision from clustering only based on the loss of Eq.10. *"No attn pretrain"* does not pre-train the attention network with our created binary mask. We observe that removing any of these components, or using an architecture based on P2PNet, leads to a noticeable performance drop. In particularly, the attention module has a significant influence on the performance of our method.

(b) *Connectivity prediction ablation study.* Table 4 presents evaluation of alternative choices for our BoneNet. In these experiments, we examine the performance of the connectivity module when it is given as input the reference joints instead of the predicted ones. In this manner, we specifically evaluate the design choices for the connectivity stage i.e., our evaluation here is not affected from any wrong predictions of the joint detection stage. Here, we report the binary classification accuracy ("Class. Acc.") i.e., whether the prediction to connect each pair of given joints agrees with the ground-truth connectivity. We also report edit distance (ED) and bone-to-bone Chamfer distance (CD-B2B), since these measures are specific to bone evaluation. We first show the performance when the MST connects joints based on Euclidean distance as cost (see "Euclidean edge cost"). We also evaluate the effect of using only the bone descriptor without the skeleton geometry encoding ($\mathbf{g}_t$) and without shape encoding ($\mathbf{g}_s$) (see "bone descriptor only", and Eq.7). We also evaluate the effect of using the bone descriptor with the skeleton geometry encoding but without shape encoding (see "bone descriptor+skel. geometry"). The best performance is achieved when all three shape, skeleton, and bone representations are used as input to BoneNet. We also observed the same trend in RootNet, where we evaluate the accuracy of predicting the root joint correctly. Skipping the skeleton geometry and shape encoding results in accuracy of 67.8%. Adding the skeleton encoding increases it to 86.8%. Using all three shape, skeleton, and joint representations achieves the best accuracy of 88.9%.

(c) *Skinning prediction ablation study*. Table 5 presents the case of removing the volumetric geodesic distance feature from input to our skinning prediction network. We observe a noticeable performance drop without it. Still, it is interesting to see that even without it, our method is better than competing methods (Table 2). We also experimented with different choices of K i.e., the number of closest bones used in our skinning prediction. Fig.12 shows the average L1-norm difference of skinning weights for $K = 1...10$ in our test set. Lowest error is achieved when $K = 5$ (we noticed the same behavior and minimum in our validation split).

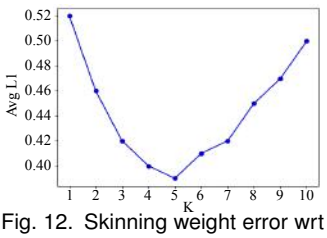

Fig. 12. Skinning weight error wrt different number K of closest bones used in our network.

7 LIMITATIONS AND CONCLUSION

We presented a method that automatically rigs input 3D character models. To the best of our knowledge, our method represents a first step towards a learning-based, complete solution to character rigging, including skeleton creation and skin weight prediction. We believe that our method is practical in various scenarios. First, we believe that our method is useful for casual users or novices, who might not have the training or expertise to deal with modeling and rigging interfaces. Another motivation for using our method is the widespread effort for democratization of 3D content creation and animation that we currently observe in online asset libraries provided with modern game engines (e.g., Unity). We see our approach as such one step towards further democratization of character animation. Another scenario of use for our method is when a large collection of 3D characters need to be rigged. Processing every single model manually would be cumbersome even for experienced artists.

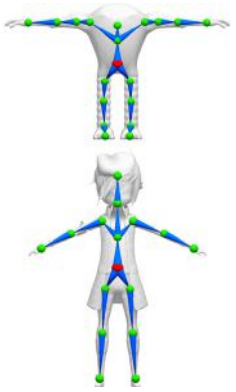

Fig. 13. Failure cases. *(Top:)* extra joints in the arms. *(Bottom:)* missing helper joints for clothes.

Our approach does have limitations, and exciting avenues for future work. First, our method currently uses a per-stage training approach. Ideally, the skinning loss could be back-propagated to all stages of the network to improve joint prediction. However, this implies differentiating volumetric geodesic distances and skeletal structure estimation, which are hard tasks. Although we trained our method such that it is more robust to different vertex sampling and tessellations, invariance to mesh resolution and connectivity is not guaranteed. Investigating the performance of other mesh neural networks (e.g., spectral) here, could be impactful. There are few cases where our method produces undesirable effects, such as putting extra arm joints (Figure 13, top). Our dataset also has limitations. It contains one rig per model. Many rigs often do not include bones for small parts, like feet, fingers, clothing and accessories, which makes our trained model less predictive of these joints (Figure 13, bottom). Enriching the dataset with more rigs could improve performance, though it might make the mapping more multi-modal than it is at present. A multi-resolution approach that refines the skeleton in a coarse-to-fine manner may instead be fruitful. Our current bandwidth parameter explores one mode of variation. Exploring a richer space to interactively control skeletal morphology and resolution is another interesting research direction. Finally, it would also be interesting to extend our method to handle skeleton extraction for point cloud recognition or reconstruction tasks.

ACKNOWLEDGMENTS

This research is partially funded by NSF (EAGER-1942069) and NSERC. Our experiments were performed in the UMass GPU cluster obtained under the Collaborative Fund managed by the Massachusetts Technology Collaborative. We thank Gopal Sharma, Difan Liu, and Olga Vesselova for their help and valuable suggestions. We also thank anonymous reviewers for their feedback.

REFERENCES

Nina Amenta and Marshall Bern. 1998. Surface Reconstruction by Voronoi Filtering. In *Proc. Symposium on Computational Geometry*.

Dominique Attali and Annick Montanvert. 1997. Computing and Simplifying 2D and 3D Continuous Skeletons. *Comput. Vis. Image Underst.* 67, 3 (1997).

Oscar Kin-Chung Au, Chiew-Lan Tai, Hung-Kuo Chu, Daniel Cohen-Or, and Tong-Yee Lee. 2008. Skeleton Extraction by Mesh Contraction. *ACM Trans. on Graphics* 27, 3 (2008).

Autodesk. 2019. *Maya, version*. www.autodesk.com/products/autodesk-maya/.

Stephen W. Bailey, Dave Otte, Paul Dilorenzo, and James F. O'Brien. 2018. Fast and Deep Deformation Approximations. *ACM Trans. on Graphics* 37, 4 (2018).

Seungbae Bang and Sung-Hee Lee. 2018. Spline Interface for Intuitive Skinning Weight Editing. *ACM Trans. on Graphics* 37, 5 (2018).

Ilya Baran and Jovan Popović. 2007. Automatic Rigging and Animation of 3D Characters. *ACM Trans. on Graphics* 26, 3 (2007).

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. 2016. Interaction networks for learning about objects, relations and physics. In *Proc. NIPS*.

Harry Blum. 1973. Biological shape and visual science (part I). *Journal of Theoretical Biology* 38, 2 (1973).

Davide Boscaini, Jonathan Masci, Emanuele Rodolà, and Michael M. Bronstein. 2016. Learning Shape Correspondence with Anisotropic Convolutional Neural Networks. In *Proc. NIPS*.

M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. 2017. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine* 34, 4 (2017).

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2014. Spectral Networks and Locally Connected Networks on Graphs. In *Proc. ICLR*.

J. Cao, A. Tagliasacchi, M. Olson, H. Zhang, and Z. Su. 2010. Point Cloud Skeletons via Laplacian Based Contraction. In *Proc. SMI*.

Yizong Cheng. 1995. Mean Shift, Mode Seeking, and Clustering. *IEEE Trans. Pat. Ana. & Mach. Int.* 17, 8 (1995).

Keenan Crane, Clarisse Weischedel, and Max Wardetzky. 2013. Geodesics in Heat: A New Approach to Computing Distance Based on Heat Flow. *ACM Trans. on Graphics* 32, 5 (2013).

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. *arXiv:1606.09375* (2016).

Sven J. Dickinson, Ales Leonardis, Bernt Schiele, and Michael J. Tarr. 2009. *Object Categorization: Computer and Human Vision Perspectives*.

Olivier Dionne and Martin de Lasa. 2013. Geodesic Voxel Binding for Production Character Meshes. In *Proc. SCA*.

O. Dionne and M. de Lasa. 2014. Geodesic Binding for Degenerate Character Geometry Using Sparse Voxelization. *IEEE Trans. Vis. & Comp. Graphics* 20, 10 (2014).

Liuhao Ge, Zhou Ren, and Junsong Yuan. 2018. Point-to-Point Regression PointNet for 3D Hand Pose Estimation. In *Proc. ECCV*.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017a. Inductive Representation Learning on Large Graphs. In *Proc. NIPS*.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017b. Representation Learning on Graphs: Methods and Applications. *IEEE Data Eng. Bull.* 40, 3 (2017).

Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. 2019. MeshCNN: A Network with an Edge. *ACM Trans. on Graphics* 38, 4 (2019).

Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, and Fei-Fei Li. 2016. Towards Viewpoint Invariant 3D Human Pose Estimation. In *Proc. ECCV*.

Mikael Henaff, Joan Bruna, and Yann LeCun. 2015. Deep Convolutional Networks on Graph-Structured Data. *arXiv:1506.05163* (2015).

Masaki Hilaga, Yoshihisa Shinagawa, Taku Kohmura, and Tosiyasu L. Kunii. 2001. Topology Matching for Fully Automatic Similarity Estimation of 3D Shapes. In *Proc. ACM SIGGRAPH*.

Fuyang Huang, Ailing Zeng, Minhao Liu, Jing Qin, and Qiang Xu. 2018. Structure-Aware 3D Hourglass Network for Hand Pose Estimation from Single Depth Image. In *Proc. BMVC*.

Hui Huang, Shihao Wu, Daniel Cohen-Or, Minglun Gong, Hao Zhang, Guiqing Li, and Baoquan Chen. 2013. L1-medial Skeleton of Point Cloud. *ACM Trans. on Graphics* 32, 4 (2013).

Alec Jacobson, Ilya Baran, Jovan Popoviundefined, and Olga Sorkine. 2011. Bounded Biharmonic Weights for Real-Time Deformation. *ACM Trans. on Graphics* 30, 4 (2011).

Alec Jacobson, Daniele Panozzo, et al. 2018. libigl: A simple C++ geometry processing library. https://libigl.github.io/.

Doug L. James and Christopher D. Twigg. 2005. Skinning Mesh Animations. *ACM Trans. on Graphics* (2005).

Timothy Jeruzalski, Boyang Deng, Mohammad Norouzi, JP Lewis, Geoffrey Hinton, and Andrea Tagliasacchi. 2019. NASA: Neural Articulated Shape Approximation. *arXiv:1912.03207* (2019).

Sagi Katz and Ayellet Tal. 2003. Hierarchical Mesh Decomposition Using Fuzzy Clustering and Cuts. *ACM Trans. on Graphics* 22, 3 (2003).

Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. 2007. Skinning with Dual Quaternions. In *Proc. I3D*.

Ladislav Kavan and Olga Sorkine. 2012. Elasticity-Inspired Deformers for Character Articulation. *ACM Trans. on Graphics* 31, 6 (2012).

Ladislav Kavan and Jiří Žára. 2005. Spherical Blend Skinning: A Real-Time Deformation of Articulated Models. In *Proc. I3D*.

Meekyoung Kim, Gerard Pons-Moll, Sergi Pujades, Seungbae Bang, Jinwook Kim, Michael J. Black, and Sung-Hee Lee. 2017. Data-Driven Physics for Human Soft Tissue Animation. *ACM Trans. on Graphics* 36, 4 (2017).

Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907* (2016).

Martin Komaritzan and Mario Botsch. 2018. Projective Skinning. *Proc. ACM Comput. Graph. Interact. Tech.* 1, 1.

Martin Komaritzan and Mario Botsch. 2019. Fast Projective Skinning. In *Proc. MIG*.

Shu Kong and Charless Fowlkes. 2018. Recurrent Pixel Embedding for Instance Grouping. In *Proc. CVPR*.

Binh Huy Le and Zhigang Deng. 2014. Robust and Accurate Skeletal Rigging from Mesh Sequences. *ACM Trans. on Graphics* 33, 4 (2014).

Binh Huy Le and Jessica K. Hodgins. 2016. Real-Time Skeletal Skinning with Optimized Centers of Rotation. *ACM Trans. on Graphics* 35, 4 (2016).

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2016. Gated graph sequence neural networks. *Proc. ICLR*.

Lijuan Liu, Youyi Zheng, Di Tang, Yi Yuan, Changjie Fan, and Kun Zhou. 2019. NeuroSkinning: Automatic Skin Binding for Production Characters with Deep Graph Networks. *ACM Trans. on Graphics* (2019).

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

Ran Luo, Tianjia Shao, Huamin Wang, Weiwei Xu, Kun Zhou, and Yin Yang. 2018. DeepWarp: DNN-based Nonlinear Deformation. *IEEE Trans. Vis. & Comp. Graphics* (2018).

N. Magnenat-Thalmann, R. Laperrière, and D. Thalmann. 1988. Joint-dependent Local Deformations for Hand Animation and Object Grasping. In *Proc. Graphics Interface '88*.

D.N. Marr and H Keith Nishihara. 1978. Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes. *Royal Society of London. Series B, Containing papers of a Biological character* 200 (1978).

Jonathan Masci, Davide Boscaini, Michael M. Bronstein, and Pierre Vandergheynst. 2015. Geodesic convolutional neural networks on Riemannian manifolds. In *Proc. ICCV Workshops*.

Models-Resource. 2019. The Models-Resource, https://www.models-resource.com/.

Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M. Bronstein. 2017. Geometric deep learning on graphs and manifolds using mixture model CNNs. In *Proc. CVPR*.

Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. 2018. V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation From a Single Depth Map. In *Proc. CVPR*.

Tomohiko Mukai and Shigeru Kuriyama. 2016. Efficient Dynamic Skinning with Low-Rank Helper Bone Controllers. *ACM Trans. on Graphics* 35, 4 (2016).

Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked Hourglass Networks for Human Pose Estimation. In *Proc. ECCV*.

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. 2017. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. In *Proc. CVPR*.

R. C. Prim. 1957. Shortest Connection Networks and some Generalizations. *The Bell Systems Technical Journal* 36, 6 (1957).

Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *Proc. NIPS*.

Yi-Ling Qiao, Lin Gao, Yu-Kun Lai, and Shihong Xia. 2018. Learning Bidirectional LSTM Networks for Synthesizing 3D Mesh Animation Sequences. *arXiv:1810.02042* (2018).

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *IEEE Trans. on Neural Networks* 20, 1 (2009).

Lior Shapira, Ariel Shamir, and Daniel Cohen-Or. 2008. Consistent Mesh Partitioning and Skeletonisation Using the Shape Diameter Function. *Visual Computer* 24, 4 (2008).

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. 2011. Real-time human pose recognition in parts from single depth images. In *Proc. CVPR*.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training region-based object detectors with online hard example mining. In *Proc. CVPR*.

Weiguang Si, Sung-Hee Lee, Eftychios Sifakis, and Demetri Terzopoulos. 2015. Realistic Biomechanical Simulation and Control of Human Swimming. *ACM Trans. on*

Graphics 34, 1 (2015).
Kaleem Siddiqi and Stephen Pizer. 2008. *Medial Representations: Mathematics, Algorithms and Applications* (1st ed.). Springer Publishing Company, Incorporated.
Kaleem Siddiqi, Ali Shokoufandeh, Sven J. Dickinson, and Steven W. Zucker. 1999. Shock Graphs and Shape Matching. *Int. J. Comp. Vis.* 35, 1 (1999).
Karan Singh and Eugene Fiume. 1998. Wires: a geometric deformation technique. In *Proc. ACM SIGGRAPH*.
Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. 2014. Earth Mover's Distances on Discrete Surfaces. *ACM Trans. on Graphics* 33, 4 (2014).
Andrea Tagliasacchi, Thomas Delame, Michela Spagnuolo, Nina Amenta, and Alexandru Telea. 2016. 3D Skeletons: A State-of-the-Art Report. *Computer Graphics Forum* (2016).
Andrea Tagliasacchi, Hao Zhang, and Daniel Cohen-Or. 2009. Curve Skeleton Extraction from Incomplete Point Cloud. *ACM Trans. on Graphics* 28, 3 (2009).
Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. 2018. Dense 3D Regression for Hand Pose Estimation. In *Proc. CVPR*.
Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. 2019. Dynamic Graph CNN for Learning on Point Clouds. *ACM Trans. on Graphics* (2019).
Rich Wareham and Joan Lasenby. 2008. Bone Glow: An Improved Method for the Assignment of Weights for Mesh Deformation. In *Proc. the 5th International Conference on Articulated Motion and Deformable Objects*.
Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2019. A Comprehensive Survey on Graph Neural Networks. *arXiv:1901.00596* (2019).
Chi Xu, Lakshmi Narasimhan Govindarajan, Yu Zhang, and Li Cheng. 2017. Lie-X: Depth Image Based Articulated Object Pose Estimation, Tracking, and Action Recognition on Lie Groups. *Int. J. Comp. Vis.* 123, 3 (2017).
Zhan Xu, Yang Zhou, Evangelos Kalogerakis, and Karan Singh. 2019. Predicting Animation Skeletons for 3D Articulated Models via Volumetric Nets. In *Proc. 3DV*.
Li Yi, Hao Su, Xingwen Guo, and Leonidas Guibas. 2017. SyncSpecCNN: Synchronized spectral CNN for 3D shape segmentation. In *Proc. CVPR*.
Kangxue Yin, Hui Huang, Daniel Cohen-Or, and Hao Zhang. 2018. P2P-NET: Bidirectional Point Displacement Net for Shape Transform. *ACM Trans. on Graphics* 37, 4 (2018).
Song Zhu and A Yuille. 1996. FORMS: A Flexible Object Recognition and Modeling System. *Int. J. Comp. Vis.* 20 (1996).

A APPENDIX: ARCHITECTURE DETAILS

Table 6 lists the layer used in each stage of our architecture along with the size of its output map. We also note that our project page with source code, datasets, and supplementary video is available at: https://zhan-xu.github.io/rig-net.

Joint Prediction Stage		
Layers	Input	Output
GMEdgeConv	$V \times 3$ (x_0)	$V \times 64$ (x_1)
GMEdgeConv	$V \times 64$	$V \times 256$ (x_2)
GMEdgeConv	$V \times 256$	$V \times 512$ (x_3)
concat(x_1, x_2, x_3)		$V \times 832$
MLP ([832, 1024])	$V \times 832$	$V \times 1024$
max_pooling & tilt	$V \times 1024$	$V \times 1024$ (x_glb)
concat$(x_0, x_1, x_2, x_3, x_glb)$		$V \times 1859$
MLP([1859, 1024, 256, 3])	$V \times 1859$	$V \times 3$
Connectivity Stage		
GMEdgeConv	$V \times 3$ (x_0)	$V \times 64$ (x_1)
GMEdgeConv	$V \times 64$	$V \times 128$ (x_2)
GMEdgeConv	$V \times 128$	$V \times 256$ (x_3)
concat(x_1, x_2, x_3)		$V \times 448$
MLP ([448, 512, 256, 128])	$V \times 448$	$V \times 128$
max_pooling & tile	$V \times 128$	$P \times 128$ (g_s)
MLP([3, 64, 128, 1024])	$K \times 3$	$K \times 1024$
max_pooling & tilt	$K \times 1024$	$P \times 1024$
MLP([1024, 256, 128])	$P \times 1024$	$P \times 128$ (g_t)
MLP([8, 32, 64, 128, 256]))	$P \times 8$	$P \times 256$ (f_ij)
concat(g_s, g_t, f_ij)		$P \times 512$
MLP([512, 128, 32, 1])	$P \times 512$	$P \times 1$
Skinning Stage		
MLP([38, 128, 64])	$V \times 38$	$V \times 64$ (x_0)
GMEdgeConv	$V \times 64$	$V \times 512$ (x_1)
max_pooling & tilt	$V \times 512$	$V \times 512$
MLP([512, 512, 1024])	$V \times 512$	$V \times 1024$ (x_glb)
GMEdgeConv	$V \times 512$ (x_1)	$V \times 256$ (x_2)
GMEdgeConv	$V \times 256$ (x_2)	$V \times 256$ (x_3)
concat(x_glb, x_3)		$V \times 1280$
MLP([1280, 1024, 512, 5])	$V \times 1280$	$V \times 5$

Table 6. RigNet architecture details. V is the number of vertices from the input mesh. K is the number of predicted joints. P is the number of candidate bones defined by all pairs of predicted joints.