Neural Marionette: Unsupervised Learning of Motion Skeleton and Latent Dynamics from Volumetric Video

Jinseok Bae, Hojun Jang, Cheol-Hui Min, Hyungun Choi, Young Min Kim

Dept. of Electrical and Computer Engineering, Seoul National University
{capoo95, j12040208, mch5048, woody101, youngmin.kim}@snu.ac.kr

Abstract

We present Neural Marionette, an unsupervised approach that discovers the skeletal structure from a dynamic sequence and learns to generate diverse motions that are consistent with the observed motion dynamics. Given a video stream of point cloud observation of an articulated body under arbitrary motion, our approach discovers the unknown low-dimensional skeletal relationship that can effectively represent the movement. Then the discovered structure is utilized to encode the motion priors of dynamic sequences in a latent structure, which can be decoded to the relative joint rotations to represent the full skeletal motion. Our approach works without any prior knowledge of the underlying motion or skeletal structure, and we demonstrate that the discovered structure is even comparable to the hand-labeled ground truth skeleton in representing a 4D sequence of motion. The skeletal structure embeds the general semantics of possible motion space that can generate motions for diverse scenarios. We verify that the learned motion prior is generalizable to the multi-modal sequence generation, interpolation of two poses, and motion retargeting to a different skeletal structure.

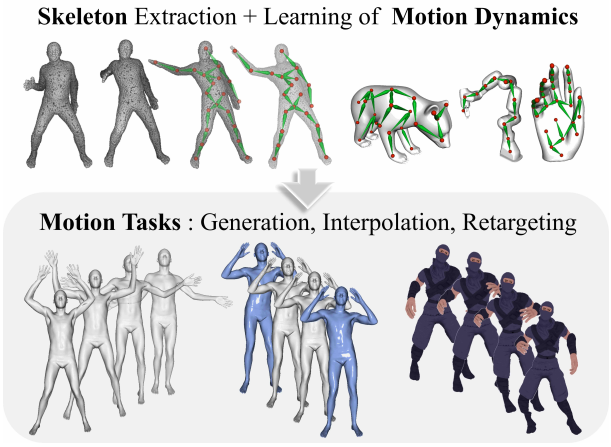

Figure 1: Overview of Neural Marionette. Our model learns to discover adaptive skeleton from volumetric video, and finds rules of motion by observing movements of skeletons.

Introduction

The skeletal structure of an articulated body (Ceccarelli 2004) has been widely deployed for robotics control (Veerapaneni et al. 2020; Ha, Xu, and Song 2020) or character animations (Xu et al. 2019b; Liu et al. 2019; Yang et al. 2020). The low-dimensional motion structure can act as an important cue to detect accurate movement and provide interaction between a human and an intelligent agent in a complex environment. Successful applications usually rely on strong priors such as human body joints, hands, or faces (Zuffi and Black 2015; Pavlakos et al. 2019; Zimmermann, Argus, and Brox 2021; Schmidtke et al. 2021) incorporated with the recent deep learning architecture. However, it is challenging to obtain the accurate structure of an unknown subject from raw observation. Some works extract skeleton using geometric priors, such as medial axis transform (Lin et al. 2021) or low-dimensional primitives (Paschalidou, Ulusoy, and Geiger 2019; Paschalidou et al. 2021), while others discover the unknown motion prior for 4D tracking or motion prediction in temporally dense observation (Bozic et al. 2021; Li et al. 2021b; Lin et al. 2020). But they are not designed to understand the motion semantics that can cover a large variation of plausible motion of a subject with an unknown skeletal structure.

In this paper, we propose Neural Marionette, a fully unsupervised framework that discovers a semantically consistent skeleton from a 3D motion sequence and learns the general motion dynamics of the discovered structure, which is illustrated in Figure 1. Our framework consists of two main stages: a *skeleton module* that determines explicit skeleton for motion, and a *dynamics module* that learns to propagate the skeletons along the time axis. Given point cloud sequence capturing the dynamic movement of an articulated body, our proposed model first learns to detect a skeleton tree without any prior knowledge of the topology. To detect candidate nodes of a skeletal graph, we adapt the keypoint detectors, which have been demonstrated to be simple yet powerful in reasoning motions from video (Minderer et al. 2019; Li et al. 2020; Suwajanakorn et al. 2018; Chen, Abbeel, and Pathak 2021). Our work extends the unsupervised keypoint detection and explicitly models the parent-child relationship between the detected keypoints to build a skeletal tree, which serves as a powerful prior of structure

to represent the motion dynamics in the subsequent stage. Given the discovered topological graph of the skeleton, the dynamics module is formulated with recurrent neural networks to embed the sequence of motion into stochastic latent variables. Specifically, the latent variable encodes the local rotation of each joint rather than location, so that general motion priors are effectively captured from different skeletons. The embedding is completely task-agnostic, and can generate a sequence of skeletal motion for any downstream tasks without further adaptation.

We demonstrate that our model can extract skeletons of various topologies including full-body humans, robots, hands, and animals. Then we evaluate the performance of the skeletal structure in motion reconstruction. Interestingly, we find that the structure discovered from our model sometimes outperforms the hand-labeled ground truth skeleton in 4D tracking. The learned dynamics is verified to generate plausible motions for three different downstream tasks: motion generation, interpolation, and retargeting. To the best of our knowledge, Neural Marionette is the first work to learn skeleton and latent dynamics from sequential 3D data, which does not exploit any categorical prior knowledge nor optimize for a specific sequence to enhance performance.

Related Works

Understanding Motion

In this work, we focus on understanding the motion of the articulated body which could be represented in terms of the skeletal structure. Our approach jointly learns the motion structure (skeleton) and the possible movement (motion dynamics), and each has been investigated in the literature. The motion structure is a shared topology of the skeleton to represent a given class of bodies. If the target class is known, for example human bodies, parametric 3D models (Anguelov et al. 2005; Loper et al. 2015; Romero, Tzionas, and Black 2017; Zuffi et al. 2017) are acquired with a large amount of human annotations. Parametric models exhibit successful achievement in a variety of applications such as shape reconstruction and pose estimation. When the structure is unknown (Palafox et al. 2021), data-driven approaches can excavate structure from observations (Xu et al. 2019a, 2020; Lin et al. 2021) with self-supervised approaches that encourage consistent topology. However, the inferred structure often is prone to errors and consequently suffers from performance degradation in motion analysis compared to sophisticated templates learned from labeled data.

After the skeletal structure defines the body as a combination of locally rigid parts, there exist a set of possible joint configurations of the given skeleton to perform plausible natural motion. Given the topology of skeleton, recent studies utilize graph neural networks to learn the complex motion patterns (Guo and Choi 2019; Mao et al. 2019; Liu et al. 2020). However, they heavily rely on accurate skeleton, and the performance is usually demonstrated in human body or production characters. To our knowledge, no previous works can discover the unknown skeletal structure and its movement that can accurately generate a large class of semantic motions.

Variational Recurrent Models

We train a generative model to represent a set of plausible motions of the given graphical structure. Variational autoencoder (VAE) (Kingma and Welling 2013) builds a latent space of the data observation that follows the desired distribution and demonstrates promising results in various generative tasks of computer vision (Eslami et al. 2016; Crawford and Pineau 2019; Burgess et al. 2019; Engelcke et al. 2019).

The latent representation can be further extended to include the temporal context of sequential data like video or speech by propagating hidden states through recurrent neural networks (RNN) (Srivastava, Mansimov, and Salakhudinov 2015). Variational recurrent neural network (VRNN) (Chung et al. 2015) is the recurrent version of VAE, which models the dependency of latent variables between neighboring timesteps. A number of works (Kosiorek et al. 2018; Minderer et al. 2019; Hajiramezanali et al. 2019; Veerapaneni et al. 2020; Lin et al. 2020) demonstrated promising results of VRNN on tasks like video prediction and dynamic link prediction. Our work also temporally extends the embedding of skeletal motion using VRNN and can successfully generate the motion sequence of the discovered skeleton.

Background

Forward Kinematics

Our skeletal structure defines the motion with the forward kinematics, which we introduce here. The position of a k-th node for skeleton in a canonical pose $\mu_{c,k}$ provides offset $d_k \in \mathbb{R}^3$ from its parent

$$d_k = \mu_{c,k} - \mu_{c,parent(k)}. \tag{1}$$

The length of the displacement $\|d_k\|$ represents the length of the bone connecting the k-th node and its parent, and is preserved under any possible deformation. A new pose of the skeleton is composed of a global translation of the root node and a set of local rotations of each joint. Specifically, the chain of forward kinematics represents the joint locations as

$$\begin{gathered} \mu_k = \mu_{parent(k)} + R_k d_k \text{ where} \\ R_k = \tilde{R}_{root} \cdots \tilde{R}_{parent(k)} \tilde{R}_k = R_{parent(k)} \tilde{R}_k. \end{gathered} \tag{2}$$

Here μ_k refers to the joint position in the current pose, and $\tilde{R}_k \in \mathbb{R}^{3\times3}$ refers to the relative rotation with respect to their parents in a local coordinate. In summary, forward kinematics encodes a pose of a skeleton with a set of rotation matrices, once the skeletal structure defines the node positions at the rest pose and directed links of parent-child relationship.

Variational Recurrent Neural Network

Variational recurrent neural network represents the observations of time steps x_t with a VAE that is composed of a prior distribution $p_\theta(z_t|h_t)$ and a posterior distribution $q_\phi(z_t|x_t, h_t)$. z_t is the latent variable that encodes the observation x_t, and sampled for generation. In addition to the ordinary VAE, both prior and posterior are conditioned on the hidden state of RNN h_t. The latent variable z_t is trained

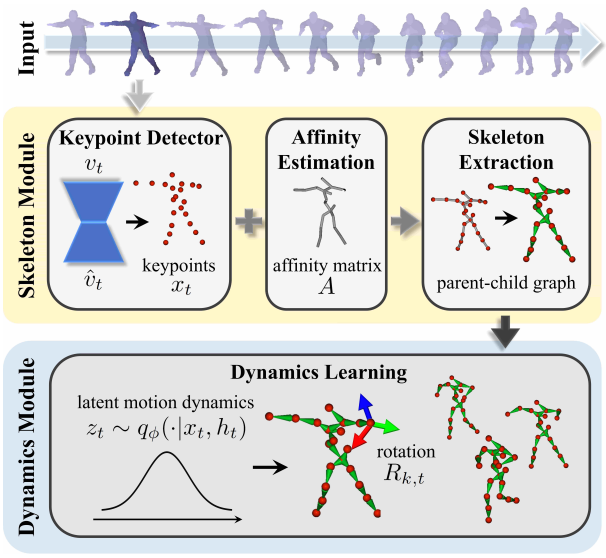

Figure 2: System pipeline of Neural Marionette. Given a voxelized sequence, skeleton module extracts skeleton, and dynamics module learns motion dynamics from the observed skeletal motion.

to maximize the likelihood of overall observation x_t with reconstructed input $\hat{x}_t$ from the decoder φ by minimizing

$$\mathcal{L}_{vrnn} = \frac{1}{T}\sum_t \|x_t - \hat{x}_t\|_2^2, \text{ where } \hat{x}_t = \varphi(z_t, h_t). \quad (3)$$

At the same time, p_θ is encouraged to track q_ϕ with the KL divergence term of $\mathcal{L}_{kl}$ to learn the distribution that matches the observation with evidence lower bound (ELBO) (Kingma and Welling 2013)

$$\mathcal{L}_{kl} = \frac{1}{T}\sum_t D_{KL}(q_\phi(z_t|x_t, h_t)\|p_\theta(z_t|h_t)). \quad (4)$$

Neural Marionette

Neural Marionette is composed of two stages: *skeleton module* that extracts the underlying skeleton, and *dynamics module* that learns the motion dynamics associated with the discovered skeleton. The overall process is described in Figure 2. Detailed structures and learning strategies of the two modules are explained below.

Skeleton Module

Given a sequence of point cloud observations, the unstructured point cloud is first discretized into binary voxels in a normalized grid $v_t \in \mathbb{R}^{G_x \times G_y \times G_z}$, $t = 1, \ldots, T$ such that we can efficiently access the neighboring observation. The skeleton module extracts the frames of keypoints from the voxelized sequences and connects the neighboring keypoints to build a shared skeleton structure for the dynamics module. The consistent skeleton acts as a light-weight and very efficient structural prior for extracting motion semantics of complex deformation.

Keypoint Detector. The keypoint detector is an encoder-decoder framework that maps the voxelized sequence v_t into the trajectory of K keypoints $\{\{x_{k,t}\}_K\}_T \in \mathbb{R}^{4\times K\times T}$. Each $x_{k,t}$ consists of a 3D location $\mu_{k,t} \in \mathbb{R}^3$ and intensity $\alpha_{k,t} \in \mathbb{R}$. Because the embedding represents the physical 3D coordinates of the keypoints, and the motion extracted with the keypoints becomes extremely interpretable in the subsequent dynamics module.

The keypoint extractor is inspired from autoencoder-based keypoint detectors on 2D video (Jakab et al. 2018; Minderer et al. 2019) and extended to 3D. The encoder first extracts grid features $c_t \in \mathbb{R}^{D\times G'_x\times G'_y\times G'_z}$ with D channels in a condensed resolution as

$$c_t = f_{feat}(v_t), \quad (5)$$

and regresses K heatmaps $m_t \in \mathbb{R}^{K\times G'_x\times G'_y\times G'_z}$

$$m_t = f_{heat}(c_t, \frac{1}{T}\sum_t v_t). \quad (6)$$

While the conventional keypoint detectors (Jakab et al. 2018; Minderer et al. 2019) do not explicitly share features between different time steps, our keypoint detector is augmented with the temporal mean of $v_{1:T}$ as shown in Eq. (6), from which the network can observe the spatio-temporal context. Each channel in m_t represents a probabilistic distribution of keypoint and we extract the keypoints $x_{k,t} = (\mu_{k,t}, \alpha_{k,t})$ from it.

The decoder is trained to recover the original sequence v_t from the keypoints x_t extracted from the encoder,

$$\begin{aligned} \hat{v}_t &= f_{dec}(g_t, g_1, c_1, v_1), \\ \text{where } g_{k,t} &= \mathcal{N}_{grid}(\mu_{k,t}, \sigma_g) \ \ \forall k. \end{aligned} \quad (7)$$

$\mathcal{N}_{grid}(\mu_{k,t}, \sigma_g)$ is Gaussian distribution discretized in a grid, with mean at $\mu_{k,t}$ and variance of hyper-parameter σ_g. Basically, the keypoint is transformed into a synthetic heatmap g_t by convolving the Gaussian kernel around the keypoint locations and reconstruct voxel difference $v_t - v_1$ with the decoding network f_{dec}. Focusing on the difference encourages the keypoints to capture dynamic area, where $v_t - v_1$ is non-zero (Minderer et al. 2019).

The encoder-decoder network is trained to find the optimal keypoints that best describes the motion of occupied voxels. Because the network encourages keypoints to capture the changing voxels, the keypoints might ignore static region. We explicitly suggest to uniformly spread the keypoints within the point cloud V_t that represents 3D coordinates p_t of occupied voxels in v_t with the volume fitting loss

$$\mathcal{L}_{vol} = \frac{1}{T}\sum_t \frac{1}{|V_t|}\sum_{p_t\in V_t} \min_k \|p_t - \mu_{k,t}\|_2^2, \quad (8)$$

that minimizes the one-directional Chamfer Distance (Fan, Su, and Guibas 2017) of V_t from keypoints x_t.

The loss function to train the autoencoder includes additional terms from previous work, namely the reconstruction loss, sparsity loss, and the separation loss. The reconstruction loss $\mathcal{L}_{recon}$ is the basic loss for an autoencoder, where

we want to best reconstruct the original voxel,

$$\mathcal{L}_{recon} = \frac{1}{T}\sum_t \text{BCE}(v_t, \hat{v}_t). \quad (9)$$

The proposed volume fitting loss complements the conventional reconstruction loss by capturing the static body parts. The remaining two loss terms are adapted from the state-of-the-art keypoint detector (Minderer et al. 2019). The sparsity loss $\mathcal{L}_{sparse}$ enforces sparsity of the heatmap,

$$\mathcal{L}_{sparse} = \frac{1}{TK}\sum_t\sum_k \|m_{k,t}\|, \quad (10)$$

and the separation loss $\mathcal{L}_{sep}$ encourages different trajectories between keypoints

$$\mathcal{L}_{sep} = \frac{1}{TK(K-1)}\sum_t\sum_k\sum_{k'\neq k} e^{-\sigma_s\|s_{k,t}-s_{k',t}\|_2^2}, \quad (11)$$

where $s_{k,t} = \mu_{k,t} - \frac{1}{T}\sum_t \mu_{k,t}$ and σ_s is a hyper-parameter.

Affinity Estimation. Along with the keypoints, our skeleton module estimates the affinity $A \in \mathbb{R}^{K\times K}$ between keypoints in order to compose edges of the skeleton. We first build decomposed affinity matrices $\{A_n \in \mathbb{R}^{K\times K}\}_N$ that focuses on $N(N < K)$ nearest neighbors of keypoints that can be combined to the final affinity matrix

$$a_{ij} = \max_n a_{n,ij}, \;\; a_{ij} \in A, a_{n,ij} \in A_n. \quad (12)$$

Our affinity estimator builds on the prior work (Bozic et al. 2021) that considers the position of the nodes as a strong prior on connections. In addition to the previously suggested losses that observe a single frame, we propose the graph trajectory loss $\mathcal{L}_{traj}$ that encourages connectivity $a_{kk'}$ between the keypoints moving in the similar path

$$\mathcal{L}_{traj} = \frac{1}{TK^2}\sum_t\sum_k \alpha_{k,t}\sum_{k'} a_{kk'}\text{C}(\mu_{k,t}, \mu_{k',t}) \quad (13)$$

given keypoint positions $\mu_{k,t}$ and intensities $\alpha_{k,t}$. $\text{C}(\cdot)$ is a function that depends on the velocities $\dot{\mu}$ and accelerations $\ddot{\mu}$ of keypoints

$$\text{C}(\mu_{k,t}, \mu_{k',t}) = \frac{1}{2} - \frac{1}{4}\left(\frac{\langle\dot{\mu}_{k,t}, \dot{\mu}_{k',t}\rangle}{\|\langle\dot{\mu}_{k,t}, \dot{\mu}_{k',t}\rangle\|} + \frac{\langle\ddot{\mu}_{k,t}, \ddot{\mu}_{k',t}\rangle}{\|\langle\ddot{\mu}_{k,t}, \ddot{\mu}_{k',t}\rangle\|}\right). \quad (14)$$

Jointly with the proposed $\mathcal{L}_{traj}$, the affinity estimator finds A_n that minimizes a loss function composed of following terms (Bozic et al. 2021): the graph local consistency loss $\mathcal{L}_{local}$, the graph time consistency loss $\mathcal{L}_{time}$, and the graph complexity loss $\mathcal{L}_{complex}$, which are

$$\mathcal{L}_{local} = \frac{1}{TK^2}\sum_t\sum_k \alpha_{k,t}\sum_{k'} a_{kk'} l_{t,kk'} \quad (15)$$

$$\mathcal{L}_{time} = \frac{1}{TK^2}\sum_t\sum_k \alpha_{k,t}\sum_{k'} a_{kk'}(l_{t,kk'} - \bar{l}_{t,kk'}) \quad (16)$$

$$\mathcal{L}_{complex} = \sum_n\sum_{n'\neq n} \|A_n \odot A_{n'}\|_F. \quad (17)$$

$\mathcal{L}_{local}$ and $\mathcal{L}_{time}$ are designed to enforce the proximity and the temporal invariance of the neighbors in Euclidean space, while $\mathcal{L}_{complex}$ helps the neighbors of each keypoint to be different by minimizing the Frobenius norm of the Hadamard product of A_n and $A_{n'}$.

Skeleton Extraction. The forward kinematics described in Eq. (2) assumes a tree structure, which starts from the root and progressively applies relative rotations on the joints of bones. After the affinity matrix is found, we choose the minimal number of edges with high affinity values that create a single connected component of keypoints. Then we find the root from the connectivity information, choosing the keypoint that has the shortest distance to all the other keypoints. Once the root is defined, we can traverse the tree and find links of parent-child relationship which can apply the forward kinematics of skeletal motion. The detailed algorithm to build the parent-child graph from the global affinity matrix is described in Sec. A.3 of the supplementary.

Dynamics Module

Using the extracted skeletal topology, the dynamics module embeds the motion into the distribution in a latent space via standard encoder structure of a variational recurrent neural network (VRNN) with Eq. (4). While the conventional VRNN learns the latent variable z_t that directly reconstructs the keypoints $\hat{x}_t$, our method encodes the local rotations of forward kinematics based on the skeletal topology.

Our decoder is implemented with the global pose decoder φ_g, and the rotation decoder φ_r. The global pose decoder φ_g decodes the translation of the root node $\hat{\mu}_{root,t}$ and the intensities $\hat{\alpha}_{k,t}$

$$\hat{\mu}_{root,t}, \{\,\hat{\alpha}_{k,t}\}_K = \varphi_g(z_t, h_t), \quad (18)$$

while the rotation decoder φ_r extracts the relative rotations

$$\{\tilde{R}_{k,t}\}_K = \varphi_r(z_t, h_t). \quad (19)$$

The positions of keypoints $\hat{\mu}_{k,t}$ are recovered from the forward kinematics process of Eq. (2), and full reconstructions $\hat{x}_t = \{(\hat{\mu}_{k,t}, \hat{\alpha}_{k,t})\}_K$ are trained to minimize Eq. (3).

We additionally propose a randomized method to mitigate the difficulty in defining the canonical relative rotation. Training with the forward kinematics requires the canonical pose of the given skeleton, which is unknown during our unsupervised setting. The canonical pose is also referred to as A-pose or T-pose, and is known a priori to define consistent relative rotations from the observation of x_t. Specifically, the canonical pose defines d_k in Eq. (1) and has to be shared for all episodes of data with the same topology of skeleton. We suggest to randomly fix the orientation of each offset $\bar{d}_k = \frac{d_k}{\|d_k\|} \in \mathbb{R}^3$ at the beginning of the training step. Then, the complete offset d_k is simply scaled from $\bar{d}_k$ by the length of a bone detected in the first frame,

$$d_k = \bar{d}_k\|\mu_{k,1} - \mu_{parent(k),1}\|_2. \quad (20)$$

The proposed randomized orientation is crucial to stabilize the training of the motion dynamics. We validate the estimated rotation in the motion retargeting task.

The dynamics module of Neural Marionette successfully embeds the motion semantics with kinematics chain used in animation or robot control, while it defines the loss in the explicit physical space. We demonstrate that our dynamics module effectively captures the distribution of motions that can generate plausible motion for various tasks.

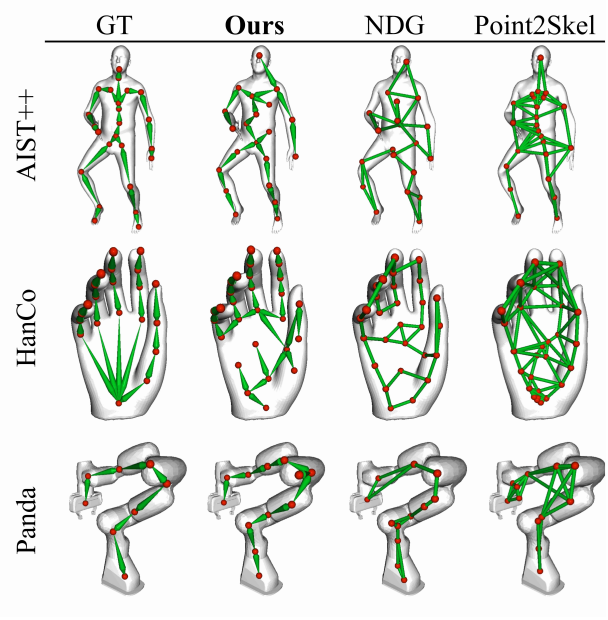

Figure 3: Qualitative comparison of extracted skeletons. Only the nodes whose intensities are above 0.2 are visualized for our model. Note that every edge has a parent-child relationship in ground truth and our model, while skeletons from NDG and Point2Skeleton do not have any hierarchy.

Experiments

Our approach extracts the skeleton of unknown topology, and we show the generalization with a wide variety of targets: **D-FAUST** (Bogo et al. 2017) and **AIST++** (Li et al. 2021a) for humans, **HanCo** (Zimmermann et al. 2019; Zimmermann, Argus, and Brox 2021) for human hands, and **Animals** (Li et al. 2021b) for various animals. We also generated a sequence of a dynamic motion of a robot arm, **Panda** with the physics-based robot simulator (Rohmer, Singh, and Freese 2013). We randomly assigned episodes in dataset into the train and test split such that the ratio of total number of frames is roughly 9:1. For quantitative evaluations, we extract a number of randomly cropped sequences for each episode in the test set, and average the results to obtain the final score of the corresponding episode.

Skeletons

We first examine the discovered skeleton topology with **AIST++**, **HanCo**, and **Panda** dataset which contain the ground truth skeleton models. We compare the quality against two recent works that find skeletons from 3D observations in a fully unsupervised manner: Point2Skeleton (Lin et al. 2021) and Neural Deformation Graph (NDG) (Bozic et al. 2021). As Point2Skeleton is a model for extracting topology in a single frame of point cloud, we additionally augment the architecture with the spatio-temporal module of CaSPR (Rempe et al. 2020) for fair comparison. With the temporal extension, both Point2Skeleton and our model can be optimized for the entire dataset to extract consistent topology. On the other hand, NDG can only be optimized for a single episode. We also would like to note that NDG uses a sequence of signed distance function grids and therefore observes richer information than point cloud.

Models	AIST++	HanCo	Panda
Ours	**0.804** (0.201)	**0.944** (0.0946)	**0.954** (0.0952)
P2S	0.755 (0.150)	0.847 (0.161)	0.937 (0.133)

Table 1: Semantic consistency score from AIST++, HanCo, and Panda dataset. Values inside parenthesis denote standard deviation for each keypoint.

Models	AIST++	HanCo	Panda
Ours	3.42 (0.924)	**1.97** (0.0932)	**1.71** (0.458)
GT	**3.11** (1.03)	2.12 (0.0971)	1.77 (0.437)

Table 2: Chamfer distance($\times 10^4$) between ground truth (GT) and reconstructed point sets that are sampled from voxel stream on 4D tracking. Values inside the parenthesis denote the 95%-confidence interval.

The recovered skeletons are compared against the ground truth skeleton in Figure 3. While the unsupervised skeleton might not exactly coincide with the ground truth, we can see that our joints are nicely spread within the volume and the links align with rigid parts of fingers or limbs. The number of nodes used for our model and baselines are 24 for **AIST++**, 28 for **HanCo**, and 12 for **Panda** dataset. Note that Neural Marionette finds the minimal skeletal graph that connects high-intensity nodes to best represent the motion. This implies that it can readily be applied to sequences with an unknown skeleton as long as the initial number of nodes is sufficient. The flexibility is an essential advantage of Neural Marionette to represent unknown motion, whereas other approaches are restricted to the fixed number of nodes. Our skeleton is deduced solely from the observation without any prior information, and therefore can be applied to various dynamic entities including humans, hands, animals, and robots as depicted in Figure 1.

To measure the quality of the recovered skeleton, we introduce semantic consistency score (SC-score). If the skeleton correctly reflects the deformation, the relative positions of nodes should be consistent with respect to the ground truth semantic labels even if the detailed topologies are different. SC-score simply indicates how consistent the closest nodes in the ground truth are for each joint in the recovered skeleton, which is represented as

$$\mathrm{SC} = \frac{1}{J} \sum_j \max_k p_j(k), \tag{21}$$

where J is the number of nodes in the ground truth, and $p_j(k)$ is the observed probability of k-th keypoint being the nearest neighbor of j-th ground truth joint. Table 1 shows that our skeleton outperforms Point2Skeleton (P2S) in every dataset and therefore reflects the correct topology and relative motion. We exclude NDG for the calculation since NDG needs to be separately optimized for every episode.

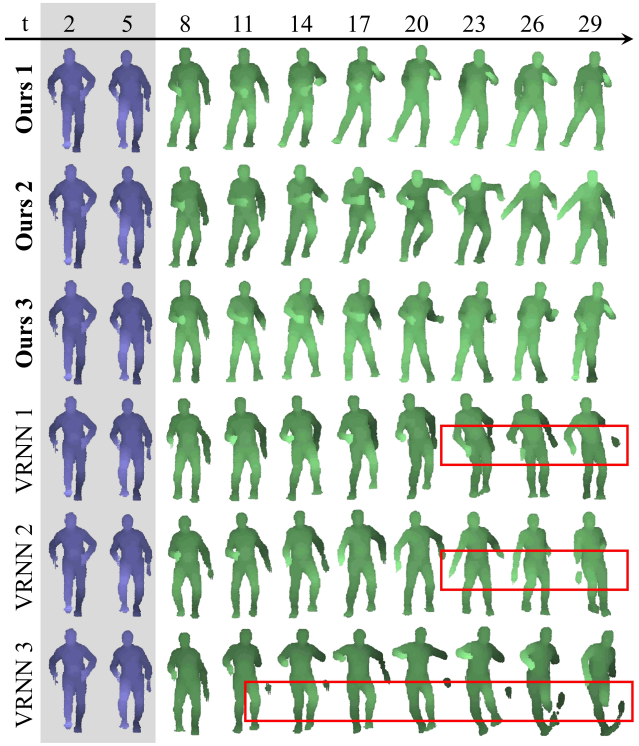

Figure 4: Result of motion generation on AIST++ dataset with five conditioned frames. Stochastic latent variables from our model can generate multi-modal future frames, which are more plausible than the results from VRNN.

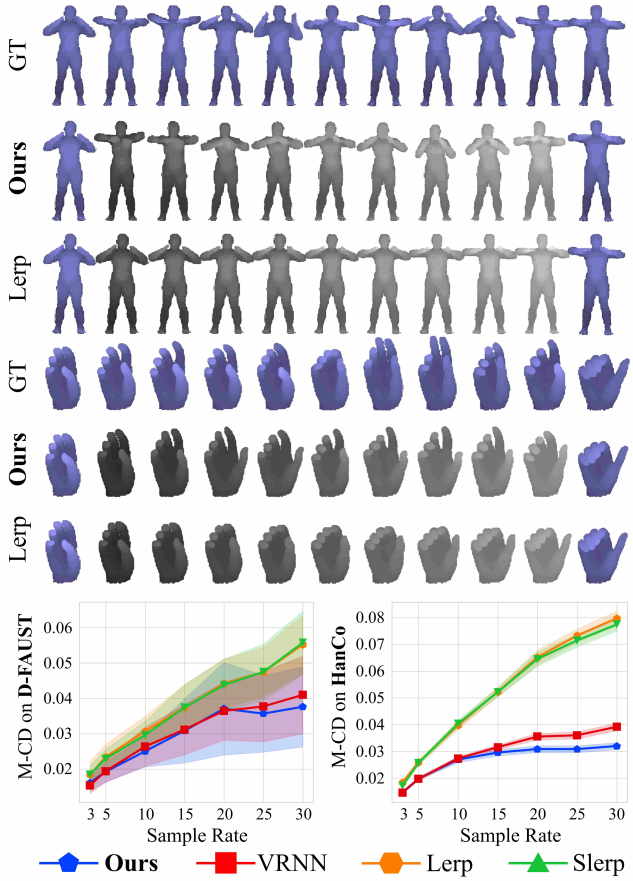

Figure 5: Result of motion interpolation on D-FAUST and HanCo dataset. Quantitative evaluations in Motion Chamfer Distance (M-CD) are plotted with 95%-confidence interval.

In addition, we observed that the hand-labeled ground truth skeletons are not necessarily optimized to represent the motion. For the baseline that is supervised with ground truth joints, we modified $\mathcal{L}_{vol}$ to minimize mean squared error between ground truth joints and the keypoints. Although the same number of nodes are used, nodes from our model show results comparable to, or sometimes even outperform the ground truth in reconstructing the given 3D sequence (Table 2). We can therefore conclude that our unsupervised skeleton can effectively capture the low-dimensional dynamics of volumetric sequence, which is further verified with various motion generation tasks.

Motion Generation

The learned dynamics of Neural Marionette can generate a variety of plausible motion sequences. Given T_{cond} frames of observation, we test how the dynamics module can predict T_{gen} frames of future sequences. We compare the quality of generated motion with the standard VRNN (Chung et al. 2015; Minderer et al. 2019) using the same input skeleton. Neural Marionette uses the tree structure and deduces the joint locations by learning the relative joint rotations, whereas VRNN directly regresses to the positions of the nodes. The qualitative comparison with **AIST++** dataset is presented in Figure 4, where T_{cond} is 5 and T_{gen} is 25. We find that our method creates much more natural and diverse motion compared to VRNN. We argue that the chain of forward kinematics of the correct skeletal representation is simple yet crucial to correctly encoding the plausible motion sequence. The individual regression of VRNN, on the other hand, results in inconsistent global positions and suffers from detached or prolonged parts. More results are available in Sec. D.2 of the supplementary material.

Motion Interpolation

We can also interpolate the starting and ending poses of keyframes, given as x_{t_s} and x_{t_e}, respectively. For motion interpolation, we generate motion as previous section without additional optimization for the different task. We sample latent variables z_t from the posterior distribution q_ϕ of the starting frame x_{t_s}, and generate N frames by sampling from prior distribution p_θ. We sample multiple trajectories of sequences, and select the one that ends with the pose closest to x_{t_e}. We can adapt a similar baseline that learns positional prior (VRNN). We also added two non-generative baselines which either linearly interpolates the joints locations of the two poses (Lerp) or spherically interpolates the local joint rotations inferred from Neural Marionette (Slerp).

Figure 5 (top) visually shows that our dynamics module more effectively interpolates the poses than baselines for

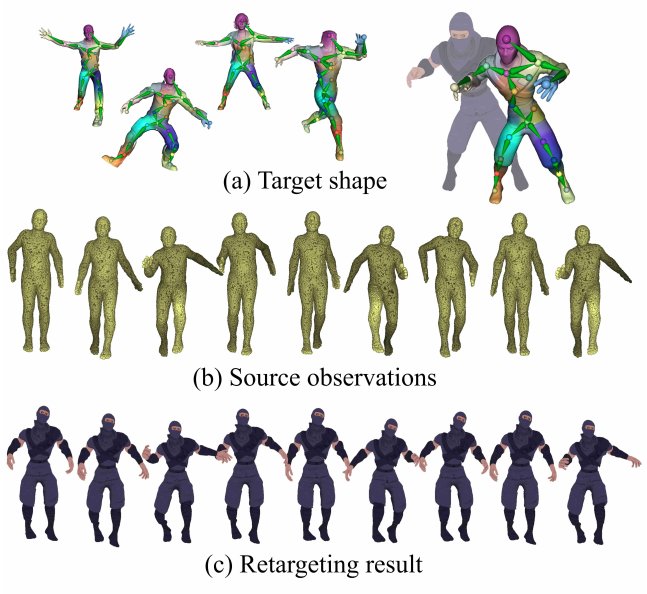

Figure 6: Overall process of motion retargeting. (a) Neural Marionette estimates the skeleton of an arbitrarily posed target shape, and (b) extracts skeletal motion from the source observations. (c) Then, the motion of the source object is retargeted into the target object.

D-FAUST and **HanCo** dataset. Sequences generated with our method better follow the ground truth motion with interesting variations in between, whereas Lerp simply moves straight to the target pose. The result with Slerp is similar to Lerp, and additional visualizations for interpolation in both datasets are contained in the supplementary material.

We suggest Motion Chamfer Distance (M-CD) using the voxel differences for quantitative comparison of reconstructed sequence, which is represented as

$$\text{M-CD} = \frac{1}{T-1}\sum_{t=2}^{T} \text{CD}(V_t^+, \hat{V}_t^+) + \text{CD}(V_t^-, \hat{V}_t^-) \quad (22)$$

where V_t^+ and V_t^- refer to sampled point clouds from the positive and negative voxels of $v_t - v_{t-1}$, and CD refers to the Chamfer Distance (Fan, Su, and Guibas 2017). When comparing the motion sequences, simply comparing the collocated occupancy can be biased toward the large overlapping static region, especially when the moving part is small with dense temporal sampling. This is because that the binary grid does not contain any correspondence information. Instead, we find the difference volume better represents motion and use it to compare with the reconstructed sequence.

The plots in Figure 5 quantitatively compare the interpolated motion against the ground truth. We can clearly see that our dynamics module outperforms all of the baselines in all of the datasets with various topologies and motions. Compared to the simple interpolation of Lerp and Slerp, the generative models of ours and VRNN performs significantly better. The result indicates that learning the motion context of joints is definitely crucial to generate plausible motion. Also our encoding with forward kinematics performs superior to directly encoding positions of keypoints with VRNN.

Motion Retargeting

We also show that the motion extracted from Neural Marionette can be transferred to a different shape with the same skeleton topology as illustrated in Figure 6. Neural Marionette encodes the relative 3D rotations of joints, enabling explicit control of motion for a given skeleton tree. We extract the source motion with pretrained Neural Marionette from **AIST++** dataset and detect the skeleton in humanoids in Mixamo dataset. We use standard linear blend skinning (LBS) and distance-based skin weights to deform the given shape, which is explained in detail in Sec. A.4 of the supplementary. Note that the ground truth skeleton and the initial canonical pose of either datasets are not known, and the poses in sequences are arbitrary. This is a highly challenging scenario compared to the standard rigging procedure, where a hand-crafted character is provided in a T-pose for adding bones and skins for further processing.

Neural Marionette deforms the mesh with the full relative transforms for forward kinematics creating natural motion. In contrast, the conventional motion representations with keypoints constrain only the locations of joints and therefore can create weird local rotations for joints. The distortion induced from rotation mismatch is prominent when the motion is retargeted to textured characters as in Figure 7.

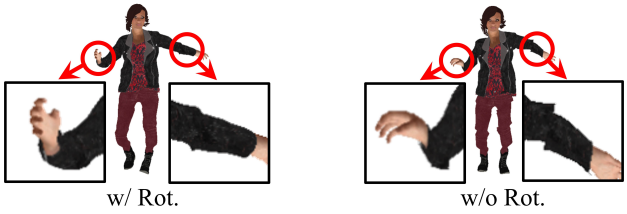

Figure 7: Comparison between motion retargeting results with and without rotations.

Conclusions

In this work, we present Neural Marionette, an unsupervised approach that captures low-dimensional motion distribution of a diverse class of unknown targets without prior information. Neural Marionette learns the motion dynamics and skeleton of a 3D model from a dynamic sequence and generates plausible motion from the learned latent distribution. Our model is explicitly designed to apply the motion using forward kinematics equipped with a skeletal tree and corresponding per-joint rotation. The direct relationship in the physical 3D space makes the representation highly interpretable. The learned distribution is readily applicable for motion generation, interpolation, and retargeting without any fine-tuning for the specific tasks. We believe that our work can be further expanded to a variety of tasks, from analyzing the movements of the unidentified target to generating plausible motions for many applications such as 3D character animation or autonomous agent control.

References

Anguelov, D.; Srinivasan, P.; Koller, D.; Thrun, S.; Rodgers, J.; and Davis, J. 2005. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, 408–416.

Bogo, F.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2017. Dynamic FAUST: Registering human bodies in motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6233–6242.

Bozic, A.; Palafox, P.; Zollhofer, M.; Thies, J.; Dai, A.; and Nießner, M. 2021. Neural Deformation Graphs for Globally-consistent Non-rigid Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1450–1459.

Burgess, C. P.; Matthey, L.; Watters, N.; Kabra, R.; Higgins, I.; Botvinick, M.; and Lerchner, A. 2019. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*.

Ceccarelli, M. 2004. *International Symposium on History of Machines and Mechanisms*. Springer.

Chen, B.; Abbeel, P.; and Pathak, D. 2021. Unsupervised Learning of Visual 3D Keypoints for Control. In *ICML*.

Chung, J.; Kastner, K.; Dinh, L.; Goel, K.; Courville, A. C.; and Bengio, Y. 2015. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28: 2980–2988.

Crawford, E.; and Pineau, J. 2019. Spatially invariant unsupervised object detection with convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3412–3420.

Engelcke, M.; Kosiorek, A. R.; Jones, O. P.; and Posner, I. 2019. Genesis: Generative scene inference and sampling with object-centric latent representations. *arXiv preprint arXiv:1907.13052*.

Eslami, S.; Heess, N.; Weber, T.; Tassa, Y.; Szepesvari, D.; Hinton, G. E.; et al. 2016. Attend, infer, repeat: Fast scene understanding with generative models. *Advances in Neural Information Processing Systems*, 29: 3225–3233.

Fan, H.; Su, H.; and Guibas, L. J. 2017. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 605–613.

Guo, X.; and Choi, J. 2019. Human motion prediction via learning local structure representations and temporal dependencies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2580–2587.

Ha, H.; Xu, J.; and Song, S. 2020. Learning a decentralized multi-arm motion planner. *arXiv preprint arXiv:2011.02608*.

Hajiramezanali, E.; Hasanzadeh, A.; Duffield, N.; Narayanan, K. R.; Zhou, M.; and Qian, X. 2019. Variational graph recurrent neural networks. *arXiv preprint arXiv:1908.09710*.

Jakab, T.; Gupta, A.; Bilen, H.; and Vedaldi, A. 2018. Unsupervised learning of object landmarks through conditional image generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 4020–4031.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kosiorek, A. R.; Kim, H.; Posner, I.; and Teh, Y. W. 2018. Sequential attend, infer, repeat: Generative modelling of moving objects. *arXiv preprint arXiv:1806.01794*.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021a. Learn to Dance with AIST++: Music Conditioned 3D Dance Generation. *arXiv preprint arXiv:2101.08779*.

Li, Y.; Takehara, H.; Taketomi, T.; Zheng, B.; and Nießner, M. 2021b. 4DComplete: Non-Rigid Motion Estimation Beyond the Observable Surface. *arXiv preprint arXiv:2105.01905*.

Li, Y.; Torralba, A.; Anandkumar, A.; Fox, D.; and Garg, A. 2020. Causal discovery in physical systems from videos. *Advances in Neural Information Processing Systems*, 33.

Lin, C.; Li, C.; Liu, Y.; Chen, N.; Choi, Y.-K.; and Wang, W. 2021. Point2Skeleton: Learning Skeletal Representations from Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4277–4286.

Lin, Z.; Wu, Y.-F.; Peri, S.; Fu, B.; Jiang, J.; and Ahn, S. 2020. Improving generative imagination in object-centric world models. In *International Conference on Machine Learning*, 6140–6149. PMLR.

Liu, L.; Zheng, Y.; Tang, D.; Yuan, Y.; Fan, C.; and Zhou, K. 2019. NeuroSkinning: Automatic skin binding for production characters with deep graph networks. *ACM Transactions on Graphics (TOG)*, 38(4): 1–12.

Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; and Ouyang, W. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 143–152.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6): 1–16.

Mao, W.; Liu, M.; Salzmann, M.; and Li, H. 2019. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9489–9497.

Minderer, M.; Sun, C.; Villegas, R.; Cole, F.; Murphy, K.; and Lee, H. 2019. Unsupervised learning of object structure and dynamics from videos. *arXiv preprint arXiv:1906.07889*.

Palafox, P.; Božič, A.; Thies, J.; Nießner, M.; and Dai, A. 2021. NPMs: Neural Parametric Models for 3D Deformable Shapes. *arXiv preprint arXiv:2104.00702*.

Paschalidou, D.; Katharopoulos, A.; Geiger, A.; and Fidler, S. 2021. Neural Parts: Learning expressive 3D shape abstractions with invertible neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3204–3215.

Paschalidou, D.; Ulusoy, A. O.; and Geiger, A. 2019. Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10344–10353.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 10975–10985.

Rempe, D.; Birdal, T.; Zhao, Y.; Gojcic, Z.; Sridhar, S.; and Guibas, L. J. 2020. Caspr: Learning canonical spatiotemporal point cloud representations. *arXiv preprint arXiv:2008.02792*.

Rohmer, E.; Singh, S. P.; and Freese, M. 2013. V-REP5: A versatile and scalable robot simulation framework. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1321–1326. IEEE.

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6): 1–17.

Schmidtke, L.; Vlontzos, A.; Ellershaw, S.; Lukens, A.; Arichi, T.; and Kainz, B. 2021. Unsupervised Human Pose Estimation through Transforming Shape Templates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2484–2494.

Srivastava, N.; Mansimov, E.; and Salakhudinov, R. 2015. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, 843–852. PMLR.

Suwajanakorn, S.; Snavely, N.; Tompson, J.; and Norouzi, M. 2018. Discovery of latent 3d keypoints via end-to-end geometric reasoning. *arXiv preprint arXiv:1807.03146*.

Veerapaneni, R.; Co-Reyes, J. D.; Chang, M.; Janner, M.; Finn, C.; Wu, J.; Tenenbaum, J.; and Levine, S. 2020. Entity abstraction in visual model-based reinforcement learning. In *Conference on Robot Learning*, 1439–1456. PMLR.

Xu, Z.; Liu, Z.; Sun, C.; Murphy, K.; Freeman, W. T.; Tenenbaum, J. B.; and Wu, J. 2019a. Unsupervised discovery of parts, structure, and dynamics. *arXiv preprint arXiv:1903.05136*.

Xu, Z.; Zhou, Y.; Kalogerakis, E.; Landreth, C.; and Singh, K. 2020. Rignet: Neural rigging for articulated characters. *arXiv preprint arXiv:2005.00559*.

Xu, Z.; Zhou, Y.; Kalogerakis, E.; and Singh, K. 2019b. Predicting animation skeletons for 3d articulated models via volumetric nets. In *2019 International Conference on 3D Vision (3DV)*, 298–307. IEEE.

Yang, Z.; Zhu, W.; Wu, W.; Qian, C.; Zhou, Q.; Zhou, B.; and Loy, C. C. 2020. Transmomo: Invariance-driven unsupervised video motion retargeting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5306–5315.

Zimmermann, C.; Argus, M.; and Brox, T. 2021. Contrastive Representation Learning for Hand Shape Estimation. *arXiv preprint arXiv:2106.04324*.

Zimmermann, C.; Ceylan, D.; Yang, J.; Russell, B.; Argus, M.; and Brox, T. 2019. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 813–822.

Zuffi, S.; and Black, M. J. 2015. The stitched puppet: A graphical model of 3d human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3537–3546.

Zuffi, S.; Kanazawa, A.; Jacobs, D. W.; and Black, M. J. 2017. 3D menagerie: Modeling the 3D shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6365–6373.