Hierarchically Structured Neural Bones for Reconstructing Animatable Objects from Casual Videos

Subin Jeon[®], In Cho[®], Minsu Kim[®], Woong Oh Cho[®], and Seon Joo Kim[®]

Yonsei University

Abstract. We propose a new framework for creating and easily manipulating 3D models of arbitrary objects using casually captured videos. Our core ingredient is a novel hierarchy deformation model, which captures motions of objects with a tree-structured bones. Our hierarchy system decomposes motions based on the granularity and reveals the correlations between parts without exploiting any prior structural knowledge. We further propose to regularize the bones to be positioned at the basis of motions, centers of parts, sufficiently covering related surfaces of the part. This is achieved by our bone occupancy function, which identifies whether a given 3D point is placed within the bone. Coupling the proposed components, our framework offers several clear advantages: (1) Users can obtain animatable 3D models of the arbitrary objects in improved quality from their casual videos, (2) users can manipulate 3D models in an intuitive manner with minimal costs, and (3) users can interactively add or delete control points as necessary. The experimental results demonstrate the efficacy of our framework on diverse instances, in reconstruction quality, interpretability and easier manipulation. Our code is available at https://github.com/subin6/HSNB.

Keywords: Animatable Model · 3D Reconstruction · Manipulation

1 Introduction

We have witnessed rapid development in creating animatable 3D models, which are playing vital roles in diverse industries, *e.g.* films, mixed reality, and games. However, such development is primarily carried out at the industry level, requiring enormous labor costs and a level of proficiency. Most of the general users, on the other hand, remain distant from this industry-level advancement, demanding more simplified ways to obtain animatable models. Recent methods [18,32,50,51] have suggested an alternative yet effective approach for general users: building animatable models from casually captured videos.

These methods employed the framework of Neural Radiance Fields (NeRF) [25] with various forms of controllable deformation models to handle the motions between frames. While a number of research [18, 32, 33, 42] adopted predefined or hand-crafted templates, *e.g.* skeletons [18, 32, 51] and 3D body models [33, 42], we stand for utilizing a set of Gaussian ellipsoids as control points,



Fig. 1: We aim to reconstruct animatable models that can be manipulated in a coarseto-fine manner, using multiple videos capturing a deformable object. The resulting 3D model can be manipulated using a hierarchical deformation model, where coarse motions are manipulated using the parent bones, and fine motions are subdivided by the child bones. We present manipulation results in novel poses.

as in BANMo [50]. These ellipsoids, so-called bones, offer a way to acquire articulated 3D models without being constrained to prior knowledge. Despite their general applicability, utilizing these bones as "control points" poses challenges in actual manipulation. This is due to the absence of structures, as these bones are distributed across the object surfaces without considering the granularity of movements, lacking correlations between bones with similar motions. Such an unstructured property also leaves room for improvement in reconstruction quality, often requiring plenty of input videos to produce plausible results.

In this paper, we present a framework for creating and easily manipulating 3D models of arbitrary objects from casual videos. We build our framework upon BANMo [50], with careful consideration to tailoring control points into well-structured forms. To provide better understanding of motions and facilitate easier manipulation of the reconstructed objects, our structured deformation model aims to decompose the motions, capture shared movements based on the granularity, and identify correlations among parts with similar motions.

To achieve this goal, we introduce a novel hierarchical bone system that represents object deformations with tree-structured bones. Our key idea is to learn the deformations in a coarse-to-fine manner: parent bones capture coarse motions of broader regions, with each child bone representing finer motion at a more specific part. We begin with a small number of bones, covering coarse parts, and gradually append child bones to cover finer motions of more specific parts. The resulting tree-structured bones identify connections between relevant bones in a fully unsupervised manner. These connections facilitate users to easily understand the structures of the motions and provide better interpretability, as well as improving reconstruction quality.

Furthermore, we suggest a regularization approach where bones are positioned at the centers of their respective parts. This is achieved using bone masks derived from the bone occupancy function and foreground masks of the objects. Instead of placing them around the surfaces as in previous methods, we extend the concept widely used in part-based generative methods [12, 30, 36] into our reconstruction pipeline for animatable models. Our bone regularization term prevents surfaces of the same part from being assigned to different bones. This facilitates our hierarchical bones to correctly capture the parts sharing motions, ensuring each bone can serve as a basis for the motions of each part.

Coupling these key ideas, the structured control points in our framework provide a more user-friendly tool for creating and manipulating 3D models with several clear advantages:

- Obtaining animatable 3D models of improved quality from casual videos.
- Manipulating 3D models in an intuitive manner with minimal effort.
- Interactively adding or deleting control points in desired parts.

We evaluate the effectiveness of our method through extensive experiments on various instances, showcasing high-quality results of the models as well as interpretable and structured control points. We also demonstrate the manipulation capability of our framework through reanimation and manipulation results.

2 Related Work

Dynamic 3D Reconstruction. Dynamic reconstruction [6,8,9,15,21,53] aims to reconstruct per-frame 3D geometry from a given video sequence. Recently, inspired by NeRF [25], its dynamic variant have significantly improved this field using only RGB videos. These dynamic methods, known as Dynamic NeRFs, can be broadly categorized into two streams. Firstly, deformation-based methods [28, 29, 34, 37] learn canonical NeRFs and per-frame deformation fields from the observation space to the canonical space simultaneously. Another line of approaches [7,10,11,19,20,44] involve learning time-conditioned NeRFs, which take time and 3D position as input and directly output color and density. Despite impressive results of such dynamic methods, the implicit learning of deformations makes it challenging to manipulate scenes into novel poses.

Animatable Object Reconstruction. Reconstructing animatable objects is a longstanding challenge in computer vision and graphics. Its goal is to reconstruct 3D models with accurate geometry that can be manipulated into novel poses. Category-specific approaches have been extensively studied with category-level templates. Model-based methods [1–3, 22, 33] represent input motions using 3D deformable models [4, 5, 23, 31, 45], while skeleton-based methods [18, 32, 35, 42, 43, 51] utilize skeletons. Recent advances in NeRF have also spurred active research in these approaches [18, 22, 32, 33, 35, 42, 51]. However, acquiring these category-specific templates necessitates either extensive 3D scan data or thorough annotation of the respective category. Such templates limits general applicability of these methods across diverse types of objects.

On the other hand, category-agnostic methods [17, 48–50] propose reconstructing animatable 3D objects from videos by learning control points simultaneously with the 3D shape, bypassing the need for predefined templates. Among these methods, BANMo [50] demonstrates promising results using a NeRF-based 3D model and linear blend skinning with implicitly learned bones. Successive researches have attempted to improve BANMo in various aspects, including root

pose decomposition [41], incomplete view coverage [38], and text-to-4D generation [40]. Orthogonal to these attempts, our method aims to improve deformation modeling to address challenge of manipulation capability, which is a crucial aspect but has received relatively little attention.

Part-level representation. Contrary to research that utilizes low-level primitives to represent motions of objects, there have been studies [12–14, 27, 30, 36] focusing on learning partial representations of 3D shapes using low-level primitives such as ellipsoids, spheres, and cubes. In these approaches, objects are composed of multiple primitives, where each primitive represents semantic parts of the object and models shapes of it. The part-based generative models are learned from a collection of data on the single class, aiming for shape abstraction [12], part understanding [13, 30], and part-based shape editing [14, 36]. We draw inspiration from this line of works to regularize our deformation parameters to be aligned with the shapes of objects, ensuring proper decomposition of motions. Furthermore, our deformation model provides hierarchical structures of primitives for motions, allowing manipulation in a coarse-to-fine manner.

3 Proposed Method

Our goal is to construct a framework for creating 3D animatable models of articulated objects from casually captured videos, offering structured bones for easier manipulation. We first deliver preliminaries [50] (Sec. 3.1), and then introduce our key components, hierarchical deformation model (Sec. 3.2), and bone occupancy function (Sec. 3.3). The overall process is outlined in Fig. 2 (a). Our method extends the overall framework of BANMo [50], with a key difference being our hierarchical deformation model and bone occupancy function.

3.1 Preliminary

BANMO [50] proposes to reconstruct animatable 3D models from RGB videos through the NeRF [25] framework. It comprises the time-invariant canonical model and the time-variant deformation model, where the deformation is defined by ellipsoidal bones and the neural skinning weight module. Given monocular RGB videos, these bones are responsible for deforming rays at each frame to the canonical pose. Then the canonical model represents the shape and the appearance of the deformed rays in the canonical pose. All components are jointly optimized together through the differentiable volume rendering.

Canonical Model represents the shape and appearance of an object as NeRFs, $g_c : (\mathbf{x}^c, \mathbf{d}) \to (\mathbf{c}, \sigma)$, which takes 3D point $\mathbf{x}^c = (x, y, z)$ in the canonical space and viewing direction $\mathbf{d} = (\phi, \theta)$ as inputs, and produces color $\mathbf{c} = (r, g, b)$ and density σ . Following VolSDF [52], the SDF value *s* is produced for mesh extraction, then *s* is transformed into σ as $\sigma = \alpha (\frac{1}{2} + \frac{1}{2} sgn(-s) (1 - exp(-\frac{|-s|}{\beta})))$, where α and β are learnable parameters.

Volume Rendering. To render a frame I_t at time t, rays r^t are cast from each pixel using a camera projection matrix. The *i*-th sampled points \mathbf{x}_i^t in r^t are



Fig. 2: (a) The overview of the proposed framework for creating 3D animatble models from videos. Each ray from the image pixel is deformed to the canonical space. Rays are deformed in a coarse-to-fine manner, using the hierarchical neural deformation model. (b) The process of hierarchical neural deformation model. Coarse motions and fine motions are composited through the bone hierarchy formulation.

deformed to the canonical space as $\mathbf{x}_i^c = T^{t \to c} \mathbf{x}_i^t$. In the canonical space, c_i and σ_i of the deformed points \mathbf{x}_i^c are queried from the canonical model. These values are composited to render the color of r^t through the volume rendering:

$$\hat{C}(r^t) = \sum_{i=1}^{N} \tau_i (1 - exp(-\sigma_i \delta_i))c_i, \qquad (1)$$

where $\tau = exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$ is the accumulated transmittance and δ_i is the distance between adjacent samples. The overall components are optimized by minimizing the differences of colors between rendered frames and given videos.

3.2 Hierarchical Neural Deformation Model

To represent motions with coarse-to-fine granularity, we introduce a hierarchical neural deformation model, as depicted in Fig. 2 (b). It takes time embedding vectors for each frame as input, and produces neural bone hierarchy for the frame. Neural bone hierarchy defines bones as Gaussian ellipsoids, with parent bones capturing coarse motions at larger regions and child bones capturing finer motions at more specific parts.

To deform a 3D point \mathbf{x}^t to canonical space, we compute poses of the leaf bones of neural bone hierarchy $\mathcal{P}^t = \{T_1^t, ..., T_B^t\}$ at time t, where $T_b^t \in SE(3)$ refers composited rigid transformation parameters through the bone hierarchy formulation for the *b*-th bone. From those parameters, the mappings between \mathcal{P}^t and the canonical poses \mathcal{P}^c are defined as

$$T_b^{t \to c} = T_b^c \cdot (T_b^t)^{-1}, \quad T_b^{c \to t} = T_b^t \cdot (T_b^c)^{-1}.$$
 (2)

Subsequently, the skinning weight $w(\mathbf{x}^t, \mathcal{P}^t)$ of \mathbf{x}^t is computed through the skinning weight module. We define the backward warping matrix $\mathcal{W}^{t\to c}_{\mathbf{x}}$ from time t to the canonical space by linear blend skinning (LBS) with w and T:

$$\mathcal{W}^{t \to c} = \sum_{b=1}^{B} w_b(\mathbf{x}^t, \mathcal{P}^t) \cdot T_b^{t \to c},\tag{3}$$



Fig. 3: Bone hierarchy diagram. Subordinate bones inherit the motion of all their parent bones (orange line). The leaf bones are used in calculating the skinning weights (blue line). Bones are gradually added during the optimization. After the optimization, users can add or delete the bones in desired regions.

where w_b is the *b*-th dimension of *w*. With the warping field, \mathbf{x}^t is deformed to the canonical space as $\mathbf{x}^c = \mathcal{W}^{t \to c} \mathbf{x}^t$. As the rigid transformation *T* is invertible, we can compute the forward warping matrix from the canonical space to time *t*:

$$\mathcal{W}^{c \to t} = \sum_{b=1}^{B} w_b(\mathbf{x}^c, \mathcal{P}^c) \cdot T_b^{c \to t}.$$
(4)

Bone Hierarchy. For a structured representation of motions, we organize neural bones in a tree-like structure, where child bones inherit the motions of their parents before making fine-grained movements. The diagram of bone hierarchy is depicted in Fig. 3. Specifically, for a specific bone at depth d, the final transformation T in the world coordinate system is composed by left-multiplying its corresponding parent transformations at previous depths in a recursive way:

$$T^{d} = \hat{T}^{1} \hat{T}^{2} \cdots \hat{T}^{d-1} \hat{T}^{d}, \tag{5}$$

where \hat{T}^d is the local transformation of the bone at depth d. Since the transformations define the center and orient of the bone, this arrangement ensures child bones are defined in the local coordinate system of their parents. Starting with a small number of bones at depth 1, as optimization proceeds, each bone is subdivided into child bones of smaller regions with finer-grained motion.

Neural Bone Representation. We follow a line of previous works [48–50] and employ 3D Gaussian ellipsoids as the primitives of our bones. Each bone consists of the rotation $R \in \mathbb{R}^{3\times 3}$, the center $\mathbf{t} \in \mathbb{R}^3$ at each time step, and a shared scale vector $\mathbf{s} \in \mathbb{R}^3$ across all time steps. These are regressed by the MLP f from the embedding vector e^t for each time t. We employ separate MLP f^d for each depth, which takes the embedding of the previous parent bone $e^{t,d}$ and the root embedding $e^{t,1}$ representing global motions. The local transformation matrix \hat{T}_i of *i*-th bone can be described as

$$\hat{T}_i^{t,1}, \mathbf{s}_i^{t,1} = f_i^1(e^{t,1}), \quad \hat{T}_i^{t,d}, \mathbf{s}_i^{t,d} = f_i^d([e^{t,1}, e^{t,d-1}]), \tag{6}$$

where f_i^d denotes the *i*-th dimension of the MLP output regressing bones at depth d, and i is a local index of the bone within its parent. The MLP f^d outputs the geometric properties of all child bones at depth d.

Skinning Weight Module. Each point \mathbf{x} is deformed by LBS with the transformation of leaf bones. The skinning weight of *b*-th leaf bone is defined as

$$w_b = \frac{exp(-d_M(\mathbf{x}, b) + \Delta w_b)}{\sum_{i=1}^{B} exp(-d_M(\mathbf{x}, b) + \Delta w_b))},$$
(7)

$$d_M(\mathbf{x}, b) = \sqrt{(\mathbf{x} - \mathbf{t}_b)^T R_b^T S_b R_b(\mathbf{x} - \mathbf{t}_b)},$$
(8)

where $d_M(\mathbf{x}, b)$ denotes the mahalanobis distance between \mathbf{x} and b-th ellipsoidal bone, and Δw_b denotes delta skinning weights computed through MLP, as in [50]. **Manipulation.** With the optimized models, users can manipulate the object into desired poses. To do this, a canonical mesh is extracted by querying the canonical model and applying the marching cube algorithm [24]. The manipulation of broad movements, which involves the motion of numerous subparts, is achieved by adjusting the parent bones, while finely-tuned motion can be easily achieved by adjusting only the sub-bones. The canonical mesh is deformed using forward warping in Eq. (4) with the new transformation parameters.

3.3 Regularizing with Bone Occupancy Function

One of the challenges in constructing a bone hierarchy lies in determining the location and the shape of the bones. Previous work [50] regularizes bone centers using Sinkhorn divergence, yet orients and scales remain under-constrained. Consequently, bones are scattered across surfaces and often larger than objects, hindering interpretability and subdivision into finer regions. To address this challenge, motivated by part-based generative methods [12, 13, 30], we propose regularization terms to align the properties of bones (center, orient, scale) with the shape of objects. The core component of our regularization is the bone occupancy function, which utilizes the mahalanobis distance $d_M(\mathbf{x}, b)$ used in the skinning weight module for identifying the occupancy.

Bone occupancy. We first model the bone occupancy function g_b , which determines the relative position with respect to the surface of bones:

$$g_b(\mathbf{x}) = d_M(\mathbf{x}, b) - \gamma, \tag{9}$$

where γ is a predefined threshold. Points inside the bone yield negative values for $g(\mathbf{x})$, while points outside the bone result in positive values. We further transform g(x) into the density function $\sigma(\frac{-g(x)}{\tau})$, which approximates 1 when \mathbf{x} is inside the bone. Here, σ is a sigmoid function, and τ is a temperature value determining the sharpness of the boundary. The bone occupancy function provides ways to relate the locations of the bones with the shapes of the objects.

Bone mask. To determine whether a 3D point \mathbf{x} is inside any bones, we define a unified bone occupancy function $G(\mathbf{x})$ by aggregating $g_b(\mathbf{x})$ of all bones:

$$G(\mathbf{x}) = \min_{b \in 1, \dots, B} g_b(\mathbf{x}). \tag{10}$$

With the density obtained from $G(\mathbf{x})$, we construct 2D bone masks M_{bone} by accumulating density values along the ray. We compute the bone mask loss by comparing them with object mask M_{GT} as

$$\mathcal{L}_{bone} = \sum ||M_{bone} - M_{GT}||^2.$$
(11)

By regularizing through the bone mask loss, we constrain the location and shape of bones to align with the actual shape of the objects.

Overlap & coverage loss. We further regularize the properties of bones based on the bone occupancy function. We extract surface points \mathcal{V} of the canonical model $g_c(\cdot)$ by applying the marching cube algorithm to the output. From the points in \mathcal{V} , we impose an overlap loss, enforcing that each point is occupied by a maximum of λ number of bones:

$$\mathcal{L}_{overlap} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{x} \in \mathcal{V}} max \Big(0, \sum_{b=1}^{B} \sigma \big(\frac{-g_b(x)}{\tau} \big) - \lambda \Big).$$
(12)

In addition, we apply a coverage loss to ensure that each bone occupies a certain portion of the entire region:

$$\mathcal{L}_{cover} = \sum_{b=1}^{B} \sum_{\mathbf{x} \in \mathcal{N}} \left(max(0, g_b(x)) \right), \tag{13}$$

where \mathcal{N} denotes the N closest points among \mathcal{V} with respect to mahalanobis distance $d_M(\mathbf{x}, b)$ to the bone.

3.4 Optimization

Our overall system is optimized on given monocular RGB videos, including 2D masks, optical flows, and dense-CSE features extracted from them. We compute the reconstruction loss term \mathcal{L}_{recon} and cycle loss term \mathcal{L}_{cycle} in BANMo [50], incorporating additional loss terms related to bones:

$$\mathcal{L} = \mathcal{L}_{recon} + \mathcal{L}_{cycle} + \mathcal{L}_{bone} + \mathcal{L}_{overlap} + \mathcal{L}_{cover}.$$
 (14)

We refer to Supplement for a more detailed description of \mathcal{L}_{recon} and \mathcal{L}_{cycle} .

Coarse-to-fine motion optimization. To optimize the hierarchical neural deformation system, we propose a coarse-to-fine motion optimization scheme. We initially optimize depth-1 bones that are responsible for coarse motion with larger region. During the optimization, we gradually add child bones to the previous bones to progressively capture fine motions.

Implementation details. In the optimization process, we start with five initial bones for animals and six for humans. After establishing the initial set of bones (parent bones), two additional bones (child bones) are added to each of the existing bones in subsequent stages. The optimization for each depth involves 20k iterations. We use two NVIDIA GeForce RTX 3090 GPUs for the optimization, and each stage takes less than 3 hours in our environment. Please refer Supplement to more implementation details.

Table 1: Quantitative results on Eagle and AMA. * indicates methods that utilize predefined skeletons for optimization. (r) indicates reproduced results.

Method	ViSER		BANMo		BANMo(r)		CAMM*		RAC*		Ours	
	CD	F2	CD	F2	CD	F2	CD	F2	CD	F2	CD	F2
Eagle	19.22	24.76	8.1	56.7	4.66	81.44	4.50	81.21	-	-	4.64	81.59
Swing	16.29	19.95	9.1	57.0	7.33	64.88	9.02	56.00	6.10	70.33	7.11	65.88
Samba	23.28	22.47	-	-	7.22	64.99	7.50	62.17	<u>6.63</u>	67.71	6.15	72.07

4 Experiment

4.1 Experimental Setup

Datasets. We evaluate our method on objects with diverse categories, including humans and animals. AMA haman dataset [39] includes multi-view videos capturing actor performances. We use Swing and Samba sequences for our evaluation on humans, and treat them as monocular videos. We also use Eagle and Cat data from BANMo dataset [50] for animals. Eagle contains videos that are rendered with an animated 3D eagle model, while Cat contains casually captured monocular videos. In the preprocessing phase, we utilize off-the-shelf models, specifically PointRend [16], VCN-robust [47], and CSE [26], to extract object masks, optical flow, and CSE features. We employ the videos of Swing, Samba, and Eagle for quantitative evaluation by comparing them to the ground-truth 3D mesh. We provide more descriptions of datasets and results of diverse animal species in Supplement.

Metrics. We evaluate the quality of reconstructed 3D objects with the following criteria. Chamfer Distance (CD) measures the average distance between the ground truth mesh and the estimated surface points. We additionally measure F-score at distance thresholds d = 2% (F2) of the longest edge of the axis-aligned object bounding box. Due to the scale ambiguity, we align the estimated 3D mesh to the ground-truth mesh using Iterative Closest Point before evaluation.

Baselines. We compare our results with both template-free methods [49,50] and skeleton-based method [17,51]. **ViSER** [49] reconstructs 3D articulated objects by learning deformation parameters guided by video-specific surface embeddings. They utilize 36 ellipsoidal bones for optimization. **BANMo** [50] estimates the pose of the objects using Gaussian ellipsoid bones with canonical NeRF. Total 25 bones are used for all categories of objects. **CAMM** [17] utilizes kinematic chains from RigNet [46] on top of BANMo to mitigate the challenges associated with manipulating Gaussian bones. Finally, **RAC** [51] reconstructs category-level 3D models. RAC uses pre-defined skeleton and learns to capture video-specific morphology from videos of diverse instances within the same category.

4.2 3D Reconstruction

Quantitative comparison. We first quantitatively evaluate the 3D reconstruction results for objects across various categories. For fair comparisons, we also



Fig. 4: Qualitative comparisons with template-free methods (ViSER, BANMo) and skeleton-based methods (CAMM, RAC). The 3D reconstruction results and the corresponding control points are described. We omit the eagle result for RAC as they require skeletons for reconstruction, which are not provided.

provide the reproduced results of BANMo as well as the original results reported in their paper. Due to the absence of the skeleton for eagles, results of RAC on Eagle are omitted. As shown in Table 1, our approach outperforms all template-free methods across all datasets. We also achieve comparable results with skeleton-based baselines without exploiting predefined structural knowledge. It is worth noting that our method achieves comparable or better results on Eagle using fewer control points compared to other baselines. Our method uses only 10 leaf bones for Eagle, whereas other baselines use 25 or more bones to represent deformation. This demonstrates the efficacy of our structured deformation model in capturing motions with reduced control points, achieving compelling results and potentially improving manipulation interfaces for users.

Qualitative comparison. Fig. 4 describes 3D reconstruction results on Samba, Cat, and Eagle datasets. Our method accurately reconstructs the 3D models with details. ViSER shows over-smoothed results with inaccurate poses, which can be attributed to their explicit meshes as shape model and the lack of the ability to aggregate multiple videos. Methods exploiting NeRF and multiple videos, on the other hand, achieve compelling reconstruction results. Methods leveraging predefined skeletons for deformations (RAC and CAMM) generally perform well in capturing poses. However, they have difficulty in accurately representing fine details of the motions which are absent in their templates, *e.g.* skirts of the Samba dataset. We provide more results of such cases in Supplement.

Control points comparison. To illustrate the interpretability of our framework, we also visualize the control points of various methods in Fig. 4. For sim-

11

	Sw	ing	San	nba	Ea	gle	Cat		
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
BANMo	29.53	0.921	30.72	0.916	31.05	0.900	28.01	0.850	
CAMM	28.04	0.912	28.87	0.907	30.44	0.894	26.47	0.830	
RAC	22.82	0.878	23.90	0.878	-	-	18.25	0.782	
Ours	30.43	0.938	31.74	0.942	32.63	0.924	28.45	0.859	

 Table 2: Quantitative comparison on neural rendering.



Fig. 5: (a) Qualitative comparison on neural rendering results. (b) Qualitative comparison of the retargeted objects.

plicity, our results only visualize leaf bones, with bones sharing parents colored in the same tint. As can be seen, our bones are aligned within the body, each of which sufficiently covers the parts of the objects. Our coarse bones capture the parts in a more broader context, *e.g.* the upper body of Samba, the wings of Eagle. Our child bones in deeper levels sub-divide these coarse parts and represent finer motions at more specific components of the objects. This can be also clearly seen in Fig. 7 (a), where bones assigned to the same parent exhibit strong correlations in movements. In contrast, the resulting control points of BANMo are scattered across the object surfaces without considering the structure and the granularity of motions, resulting in difficulty of understanding and animating the 3D models. The bone hierarchy of our system provides organized control points for the deformations, enhancing understanding of controls and a more user-friendly manipulation experience.

4.3 Neural Rendering

We compare the rendered results with NeRF-based methods. For quantitative evaluation, we measure the PSNR and SSIM scores between the rendered results and the ground-truth images. As shown in Table 2, our method outperforms all baselines across diverse categories of objects, demonstrating that our hierarchical modeling of motion enhances rendering quality as well. Fig. 5 (a) illustrates the rendering results on the Cat and Samba datasets. Evident in the detailed motion of the arm (highlighted in the blue box), our method effectively captures intricate movements, resulting in clearer RGB renderings.

Table 3: Quantitative comparison Table 4: Ablation results on the number of videos.(CD) of the retargeted objects.

				#videos	1 vid		4 vids		8 vids		
#steps	50	100	150	200	# videos	CD	F2	CD	F2	CD	F2
BANMo	2 75	2.03	1.90	1.86	CAMM	17.03	38.52	10.65	48.72	7.50	62.17
Ours	2.10	1.03	1.83	1.00	BANMo	10.28	47.70	11.34	45.20	7.22	64.99
Ours	2.15	1.35	1.05	1.15	Ours	9.92	52.29	7.05	62.34	6.15	72.07

Table 5: Quantitative ablation results on the number of depths and the regularization.

Bone reg.		No reg.		Sinkhorn		Bone occupancy function					
(#depths, #bones)		(1, 6)	(1, 24)	(1, 6)	(1, 24)	(1, 6)	(1, 12)	(1, 24)	(2, 12)	(3, 24)	
Samba	CD	7.66	6.84	8.56	7.17	7.65	7.21	7.16	6.87	6.15	
	F2	61.38	67.66	57.23	65.67	61.93	63.78	65.41	66.76	72.07	
Qin m	CD	8.96	8.37	9.60	8.39	9.27	9.37	8.83	7.74	7.11	
owing	F2	55.61	59.39	52.91	59.70	54.74	54.34	58.29	61.64	65.88	

4.4 Reanimation

We further compare the reanimation capability and effectiveness of the learned control points against BANMo [50]. To this end, we conduct optimization-based motion retargeting experiments, following a previous work [43]. Given canonical shapes and corresponding bone parameters, the objective is to retarget the pose of models to a new target pose through bone adjustments. Specifically, the transform parameters of bones are optimized to minimize CD between predicted and target shapes while preserving fixed canonical shape and skinning weights. We rig the ground truth mesh of Samba and craft a sequence of 150 frames depicting a novel motion. We also provide results with various optimization steps (per frame) to illustrate the speed at which we can achieve a target pose.

As shown in Fig. 5 (b), we achieve convincing results with fewer optimization steps, thanks to our structured property that moves larger regions with similar motion simultaneously. As the number of steps increases, the fine details of the poses are further refined. In contrast, BANMo struggles with handling large motions (e.g., seating, as in the first pose), leading to collapsed body structures. Table 3 presents a quantitative comparison of the retargeted objects. We outperform the baseline at all steps, particularly with a significant margin at a small number of steps, implying better animating capability of our method.

4.5 Manipulation

We demonstrate the capability of our method in manipulating a diverse set of objects. The core advantage of our approach is that it provides a coarse-to-fine manipulation, providing easier manipulation for users. We deliver the example results of the manipulation using our framework in Fig. 6. Thanks to our tree-structured control points, we can animate various poses with a minimal number of actions. For instance, we can animate the human and cat to sit using only depth-1 bones (coarsest level), with total 5 movements. On the other hand, the unstructured bones of BANMo necessitate independent manipulation of the



Fig. 6: Manipulation results on diverse categories of objects. The left side of each column illustrates the depth 1 bones and their corresponding skinning weights, while the right side shows the manipulated results.



Fig. 7: (a) Visualization of the hierarchically structured bones at each depth. (b) Qualitative ablation results on the bone regularization terms.

bones to make the same pose, requiring total 25 movements. In addition, as our deformation model gradually captures the coarse-to-fine structures of the motions, we can flexibly add or delete some of the bones if necessary (Fig. 3). If users want to add more control points on the tail of the cat, to better capture detailed motions of it, it can be easily achieved by appending child bones to the corresponding bone. Note that such dynamic control over the number of bones is not feasible within the framework of BANMo, as its bones lack structure, making it challenging to determine the locations of new bones. We refer to Supplement for the results of dynamic addition and deletion of the bones.

4.6 Ablation Study

Hierarchical neural deformation model. We ablate our hierarchical neural deformation model by gradually increasing the depths (#depths = 1, 2, 3). We compare this to the models without our hierarchy system, which use the same number of bones in one depth (#bones = 12, 24). As reported in Table 5, even when using the same number of bones, the model with our hierarchy system yields much improved quantitative results. This indicates that capturing coarse motions at the beginning and progressively refining fine-grained movements is more effective in optimizing motions. Such progressive procedure is more depicted in Fig. 7 (a). In the case of Samba, our system first assigns a single bone

to the entire leg. As the depth increases, this coarse bone is sub-divided into more specific parts, e.g. the calf and the foot, providing correlations between bones with similar motions.

Bone Regularization. We then conduct the ablation on the bone regularization terms. We compare our model with (1) the model without regularization (No regularization) and (2) the model optimized with Sinkhorn divergence, as in previous work [50]. To explore the effects more clearly, we compare these models without our hierarchy system (#depths = 1). We deliver the results using 6 and 24 bones, which represent an insufficient and a sufficient number of bones to capture the motions, respectively. Table 5 and Fig. 7 (b) present quantitative and qualitative results. The model regularized with our bone mask loss achieves better results compared to the Sinkhorn divergence loss. Interestingly, in some cases, the model without regularization delivers the best results. Despite its quantitative results, as shown in Fig. 7 (b), bones optimized without any regularization tend to float outside of the body, making it challenging to discern which bone is responsible for a specific part. The bones regularized with our bone regularization effectively captures motions while being more appropriately placed, achieving improvement when combined with our hierarchy system.

Number of input videos. Finally, we investigate the performance with a limited number of videos. We compare the results on Samba by using a single video (1 vid), a half number of videos (4 vids), and all videos (8 vids). As shown in Table 4, we outperform baselines in all settings. BANMo suffers from correctly reconstructing models when using fewer videos, due to the absence of structures in its control points. On the other hand, our method outperforms BANMo (8 vids) with only using a half number of videos (4 vids), demonstrating the robustness and effectiveness of our structured deformation model.

5 Discussion and conclusion

We presented a new framework for creating and animating 3D models, from a set of casually captured videos. Our hierarchy neural deformation model provides a way to acquire structured bone representations, without exploiting prior structural knowledge, thereby enabling the general applicability of our method. Combined with the regularization based on the bone occupancy function, our method facilitates easier and interpretable manipulation. Our approach alleviates the requirements for obtaining animatable models of arbitrary objects, with more comprehensive control points that truly function as "control points".

Limitation and future works. While our structured deformation model provides connections between the bones having similar movements, we expect the motions of the articulated objects can be better captured by the dynamic discovery of joints and conjunction. Moreover, extending our framework to scenes having multiple objects is a worth exploring subject, which we plan to resolve in our future research.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00124), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF- 2022R1A2C2004509), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (Artificial Intelligence Graduate School Program, Yonsei University, under Grant 2020-0-01361).

References

- Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single rgb camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1175– 1186 (2019)
- Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Detailed human avatars from monocular video. In: 2018 International Conference on 3D Vision (3DV). pp. 98–109. IEEE (2018)
- Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8387–8397 (2018)
- Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM SIGGRAPH 2005 Papers, pp. 408–416 (2005)
- 5. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 157–164 (2023)
- Bozic, A., Zollhofer, M., Theobalt, C., Nießner, M.: Deepdeform: Learning nonrigid rgb-d reconstruction with semi-supervised data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7002– 7012 (2020)
- Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
- Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Transactions on Graphics (ToG) 34(4), 1–13 (2015)
- Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S.R., Kowdle, A., Escolano, S.O., Rhemann, C., Kim, D., Taylor, J., et al.: Fusion4d: Real-time performance capture of challenging scenes. ACM Transactions on Graphics (ToG) 35(4), 1–13 (2016)
- Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12479– 12488 (2023)
- Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5712–5721 (2021)

- 16 S. Jeon et al.
- Genova, K., Cole, F., Sud, A., Sarna, A., Funkhouser, T.: Local deep implicit functions for 3d shape. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4857–4866 (2020)
- Genova, K., Cole, F., Vlasic, D., Sarna, A., Freeman, W.T., Funkhouser, T.: Learning shape templates with structured implicit functions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7154–7164 (2019)
- Hertz, A., Perel, O., Giryes, R., Sorkine-Hornung, O., Cohen-Or, D.: Spaghetti: Editing implicit shapes through part aware generation. ACM Transactions on Graphics (TOG) 41(4), 1–20 (2022)
- Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., Stamminger, M.: Volumedeform: Real-time volumetric non-rigid reconstruction. In: Computer Vision– ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. pp. 362–379. Springer (2016)
- Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9799–9808 (2020)
- Kuai, T., Karthikeyan, A., Kant, Y., Mirzaei, A., Gilitschenski, I.: Camm: Building category-agnostic and animatable 3d models from monocular videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6586–6596 (2023)
- Li, R., Tanke, J., Vo, M., Zollhöfer, M., Gall, J., Kanazawa, A., Lassner, C.: Tava: Template-free animatable volumetric actors. In: European Conference on Computer Vision. pp. 419–436. Springer (2022)
- Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., et al.: Neural 3d video synthesis from multi-view video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5521–5531 (2022)
- Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6498–6508 (2021)
- Lin, W., Zheng, C., Yong, J.H., Xu, F.: Occlusionfusion: Occlusion-aware motion estimation for real-time dynamic 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1736– 1745 (2022)
- Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. ACM transactions on graphics (TOG) 40(6), 1–16 (2021)
- Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)
- 24. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. SIGGRAPH Comput. Graph. (1987)
- Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM 65(1), 99–106 (2021)
- Neverova, N., Novotny, D., Szafraniec, M., Khalidov, V., Labatut, P., Vedaldi, A.: Continuous surface embeddings. Advances in Neural Information Processing Systems 33, 17258–17270 (2020)
- 27. Noguchi, A., Iqbal, U., Tremblay, J., Harada, T., Gallo, O.: Watch it move: Unsupervised discovery of 3d joints for re-posing of articulated objects. In: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3677-3687 (2022)

- Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5874 (2021)
- Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. ACM Trans. Graph. 40(6) (dec 2021)
- Paschalidou, D., Katharopoulos, A., Geiger, A., Fidler, S.: Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3204–3215 (2021)
- Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019)
- Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14314–14323 (2021)
- 33. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9054–9063 (2021)
- Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
- Su, S.Y., Yu, F., Zollhöfer, M., Rhodin, H.: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. Advances in Neural Information Processing Systems 34, 12278–12291 (2021)
- 36. Tertikas, K., Despoina, P., Pan, B., Park, J.J., Uy, M.A., Emiris, I., Avrithis, Y., Guibas, L.: Partnerf: Generating part-aware editable 3d shapes without 3d supervision. arXiv preprint arXiv:2303.09554 (2023)
- 37. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12959–12970 (2021)
- Tu, T., Li, M.F., Lin, C.H., Cheng, Y.C., Sun, M., Yang, M.H.: Dreamo: Articulated 3d reconstruction from a single casual video. arXiv preprint arXiv:2312.02617 (2023)
- Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. In: Acm Siggraph 2008 papers, pp. 1–9 (2008)
- 40. Wang, X., Wang, Y., Ye, J., Wang, Z., Sun, F., Liu, P., Wang, L., Sun, K., Wang, X., He, B.: Animatabledreamer: Text-guided non-rigid 3d model generation and reconstruction with canonical score distillation. arXiv preprint arXiv:2312.03795 (2023)
- Wang, Y., Dong, Y., Sun, F., Yang, X.: Root pose decomposition towards generic non-rigid 3d reconstruction with monocular videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13890–13900 (2023)
- 42. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: Humannerf: Free-viewpoint rendering of moving people from monocular video.

In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. pp. 16210–16220 (2022)

- Wu, Y., Chen, Z., Liu, S., Ren, Z., Wang, S.: Casa: Category-agnostic skeletal animal reconstruction. Advances in Neural Information Processing Systems 35, 28559–28574 (2022)
- Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9421–9431 (2021)
- 45. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10965–10974 (2019)
- Xu, Z., Zhou, Y., Kalogerakis, E., Landreth, C., Singh, K.: Rignet: Neural rigging for articulated characters. ACM Trans. on Graphics **39** (2020)
- 47. Yang, G., Ramanan, D.: Volumetric correspondence networks for optical flow. Advances in neural information processing systems **32** (2019)
- 48. Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Chang, H., Ramanan, D., Freeman, W.T., Liu, C.: Lasr: Learning articulated shape reconstruction from a monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15980–15989 (2021)
- Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Liu, C., Ramanan, D.: Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. Advances in Neural Information Processing Systems 34, 19326–19338 (2021)
- Yang, G., Vo, M., Neverova, N., Ramanan, D., Vedaldi, A., Joo, H.: Banmo: Building animatable 3d neural models from many casual videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2863– 2873 (2022)
- Yang, G., Wang, C., Reddy, N.D., Ramanan, D.: Reconstructing animatable categories from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16995–17005 (2023)
- Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. Advances in Neural Information Processing Systems 34, 4805–4815 (2021)
- 53. Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., et al.: Real-time non-rigid reconstruction using an rgb-d camera. ACM Transactions on Graphics (ToG) 33(4), 1–12 (2014)