Reconstructing Animatable Categories from Videos

Gengshan Yang Chaoyang Wang N Dinesh Reddy Deva Ramanan
Carnegie Mellon University

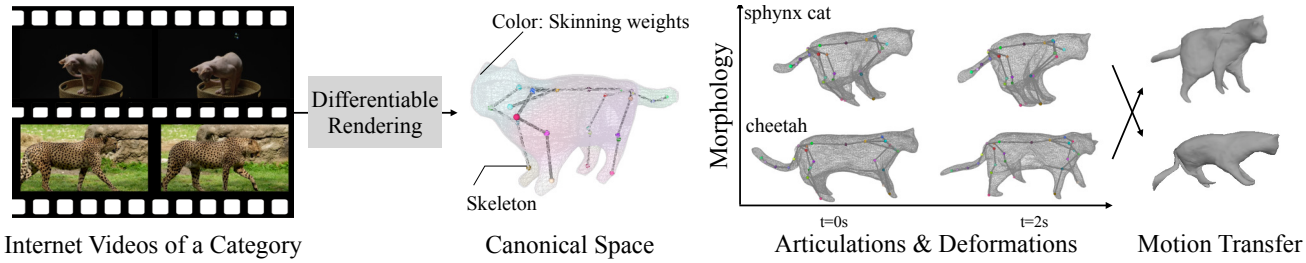

Figure 1. Given videos of a deformable category and a skeleton, we reconstruct an animatable 3D model that factorizes variations *across* instances (e.g., `cheetah`'s and `sphynx`'s are both `cats` but with different shape morphology, skeleton dimensions, and texture) from time-specific variations *within* an instance (e.g., skeleton articulations and elastic shape deformation). **Left**: Input videos; **Middle-left**: 3D shape, skeleton, and skinning weights (visualized as surface colors) in the canonical space; **Middle-right**: Disentangled between-instance and within-instance variations over time. **Right**: Morphology and motion transferred across the two instances.

Abstract

Building animatable 3D models is challenging due to the need for 3D scans, laborious registration, and rigging. Recently, differentiable rendering provides a pathway to obtain high-quality 3D models from monocular videos, but these are limited to rigid categories or single instances. We present RAC, a method to build category-level 3D models from monocular videos, disentangling variations over instances and motion over time. Three key ideas are introduced to solve this problem: (1) specializing a category-level skeleton to instances, (2) a method for latent space regularization that encourages shared structure across a category while maintaining instance details, and (3) using 3D background models to disentangle objects from the background. We build 3D models for humans, cats and dogs given monocular videos. Project page: https://gengshan-y.github.io/rac-www/.

1. Introduction

We aim to build animatable 3D models for deformable object categories. Prior work has done so for targeted categories such as people (e.g., SMPL [1, 30]) and quadruped animals (e.g., SMAL [4]), but such methods appear challenging to scale due to the need of 3D supervision and registration. Recently, test-time optimization through differentiable rendering [40, 41, 44, 56, 69] provides a pathway to generate high-quality 3D models of deformable objects and scenes from monocular videos. However, such models are typically built *independently* for each object instance or scene. In contrast, we would like to build *category* models that can generate different instances along with deformations, given *causally-captured video collections*.

Though scalable, such data is challenging to leverage in practice. One challenge is how to learn the *morphological variation* of instances within a category. For example, `huskys` and `corgis` are both `dogs`, but have different body shapes, skeleton dimensions, and texture appearance. Such variations are difficult to disentangle from the variations *within* a single instance, e.g., as a dog articulates, stretches its muscles, and even moves into different illumination conditions. Approaches for disentangling such factors require enormous efforts in capture and registration [1, 5], and doing so without explicit supervision remains an open challenge.

Another challenge arises from the impoverished nature of in-the-wild videos: objects are often *partially observable* at a limited number of viewpoints, and input signals such as segmentation masks can be inaccurate for such "in-the-wild" data. When dealing with partial or impoverished video inputs, one would want the model to listen to the common structures learned across a category – e.g., dogs have two ears. On the other hand, one would want the model to

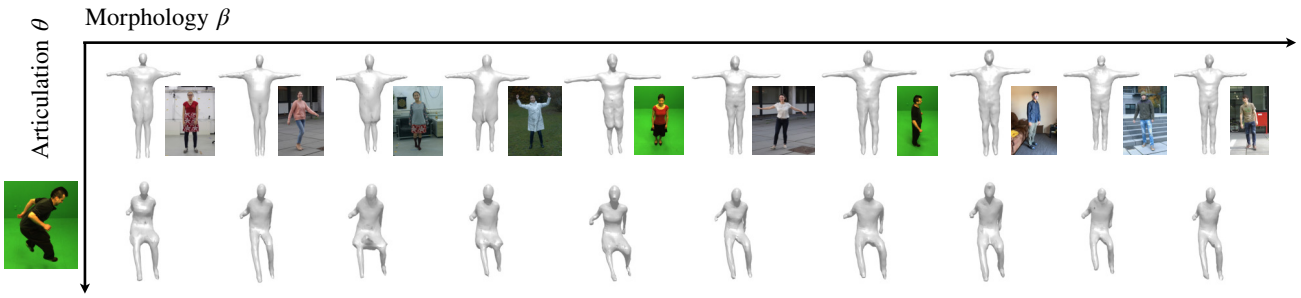

Figure 2. **Disentangling morphologies β and articulation θ.** We show different morphologies (body shape and clothing) given the same rest pose (**top**) and bouncing pose (**bottom**).

stay faithful to the input views.

Our approach addresses these challenges by exploiting three insights: (1) We learn skeletons with constant bone lengths within a video, allowing for better disentanglement of between-instance morphology and within-instance articulation. (2) We regularize the unobserved body parts to be coherent across instances while remaining faithful to the input views with a novel code-swapping technique. (3) We make use of a category-level background model that, while not 3D accurate, produces far better segmentation masks. We learn animatable 3D models of cats, dogs, and humans which outperform prior art. Because our models register different instances with a canonical skeleton, we also demonstrate motion transfer across instances.

2. Related Works

Model-based 3D Reconstruction. A large body of work in 3D human and animal reconstruction uses parametric shape models [30, 42, 55, 61, 73, 74], which are built from registered 3D scans of human or animals, and serve to recover 3D shapes given a single image or video at test time [2, 3, 3, 21, 21, 46, 72]. A recent research focus is to combine statistical human body mode with implicit functions [17, 25, 47–49, 63, 71] to improve the robustness and fidelity of the reconstruction. Although parametric body models achieve great success in reconstructing humans with large amounts of ground-truth 3D data, it is unclear how to apply the same methodology to categories with limited 3D data, such as animals, and how to scale to real-life imagery with diverse clothing and body poses. RAC builds category-level shape models from in-the-wild videos and demonstrates the potential to reconstruct 3D categories without sophisticated manual processing.

Category Reconstruction from Image Collections. A number of recent methods build deformable 3D models of object categories from images with weak 2D annotations, such as keypoints and object silhouettes, obtained from human annotators or predicted by off-the-shelf models [13, 18, 22, 28, 53, 59, 70]. However, those methods do not distinguish between morphological variations and motion over time. Moreover, they often apply heavy regularization on shape and deformation to avoid degenerate solutions, which also smooths out fine-grained details. Recent research combines neural implicit functions [33, 34] with category modeling in the context of 3D data generation [6, 7, 37], where shape and appearance variations over a category are modeled with conditional NeRFs. However, reconstructions are typically focused on near-rigid objects such as faces and vehicles.

Articulated Object Reconstruction from Videos. Compared to image collections, videos provide signals to reconstruct object motions and disentangle them from morphological variations. Some works [26, 39, 51] reconstruct articulated bodies from videos, but they either assume synchronized multi-view recordings or articulated 3D skeleton inputs that make their approaches less general. Some other works [67–69] learn animatable 3D models from monocular videos capturing the same object instance, without disentangling morphology and motion. There are recent methods [27, 58, 68] using in-the-wild videos to reconstruct 3D models animals, but their quality are relatively low.

3. Method

Given video recordings of different instances from a category and a pre-defined skeleton, we build animatable 3D models including instance-specific morphology (Sec. 3.1), time-varying articulation and deformation (Sec. 3.2), as well as a video-specific 3D background model (Sec. 3.3). The models are optimized using differentiable rendering (Sec. 3.4). An overview is shown in Fig. 3.

3.1. Between-Instance Variation

Fusing videos of different instances into a category model requires handling the morphological variations, which includes the changes in both *internal skeleton* and *outward appearance* (shape and color). We define a video-

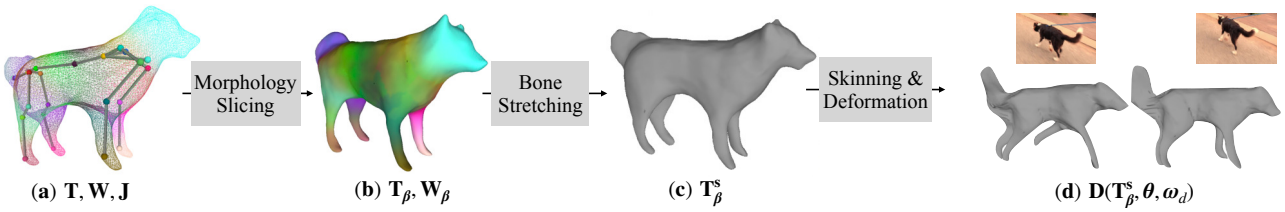

Figure 3. **Morphological variations vs time-varying articulation and deformation.** (a) Canonical shape $\mathbf{T}$, skinning weights $\mathbf{W}$, and joint locations $\mathbf{J}$. (b) To represent morphological differences between instances, we use a morphology code $\boldsymbol{\beta}$ that specifies instance shape and appearance $\mathbf{T}_{\boldsymbol{\beta}}$, skinning weights $\mathbf{W}_{\boldsymbol{\beta}}$ for a canonical skeleton $\mathbf{J}$. (c) $\boldsymbol{\beta}$ also predicts a change in bone lengths $\Delta\mathbf{J}_{\boldsymbol{\beta}}$ which further *stretches* instance shape into $\mathbf{T}^s_{\boldsymbol{\beta}}$ by elongating body parts. (d) Time-varying articulations are modeled with an articulation vector $\boldsymbol{\theta}$ by linearly blending rigid bone transformations in the dual quaternion space. Time-varying deformations (such as muscle deformation) are modeled with a deformation vector $\boldsymbol{\omega}_d$ through invertible 3D warping fields.

specific morphology code $\boldsymbol{\beta}$ to control the variations of both the shape and the skeleton.

To model between-instance shape variations, one could use dense warping fields to deform a canonical template into instance-specific shapes [62]. However, warping fields cannot explain topological changes (e.g., different clothing). Instead, we define a hierarchical representation: a conditional canonical field [8, 41, 57] to handle fine-grained variations over a category (e.g., the ears of dogs) and a stretchable bone model [14, 60] to represent coarse shape variations (e.g., height and size of body parts).

Conditional Field T. In the canonical space, a 3D point $\mathbf{X} \in \mathbb{R}^3$ is associated with three properties: signed distance $d \in \mathbb{R}$, color $\mathbf{c} \in \mathbb{R}^3$, and canonical features $\boldsymbol{\psi} \in \mathbb{R}^{16}$, which is used to register pixel observations to the canonical space [35, 68]. These properties are predicted by multi-layer perceptron (MLP) networks:

$$(d, \mathbf{c}^t) = \mathbf{MLP}_{\text{SDF}}(\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\omega}_a), \quad (1)$$

$$\boldsymbol{\psi} = \mathbf{MLP}_{\boldsymbol{\psi}}(\mathbf{X}), \quad (2)$$

where the shape and color are conditioned on a video-specific morphology code $\boldsymbol{\beta} \in \mathbb{R}^{32}$ [16, 37]. We further ask the color to be dependent on an appearance code $\boldsymbol{\omega}_a \in \mathbb{R}^{64}$ that captures frame-specific appearance such as shadows and illumination changes [32].

Skeleton J. Unlike shape and color, the bone structures are not directly observable from imagery, making it ambiguous to infer. Methods for automatic skeletal rigging [23, 38, 65] either heavily rely on shape priors, or appear sensitive to input data. Instead, we provide a category-level skeleton topology, which has a fixed tree topology with (B+1) bones and B joints (B=25 for quadruped and B=18 for human). To model cross-instance morphological changes, we define per-instance joint locations as:

$$\mathbf{J} = \mathbf{MLP}_{\mathbf{J}}(\boldsymbol{\beta}) \in \mathbb{R}^{3\times B}. \quad (3)$$

As we will discuss next, the change in joint locations not only stretches the skeleton, but also results in the elongation of canonical shapes as shown in Fig. 3 (c). The skeleton topology is fixed through optimization but $\mathbf{J}$ is specialized to each video.

Skinning Field W. For a given 3D location $\mathbf{X}$, we define skinning weight vector $\mathbf{W} \in \mathbb{R}^{B+1}$ following BANMo:

$$\mathbf{W} = \sigma_{\text{softmax}}\big(d_\sigma(\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}) + \mathbf{MLP}_{\mathbf{W}}(\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta})\big), \quad (4)$$

where $\boldsymbol{\theta}$ is a articulation code and $d_\sigma(\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta})$ is the Mahalanobis distance between $\mathbf{X}$ and Gaussian bones under articulation $\boldsymbol{\theta}$ and morphology $\boldsymbol{\beta}$, refined by a delta skinning MLP. Each Gaussian bone has three parameters for center, orientation, and scale respectively, where the centers are computed as the midpoint of two adjacent joints, the orientations are determined by the parent joints, and the scales are optimized.

Stretchable Bone Deformation. To represent variations of body dimension and part size, prior work [30, 74] learns a PCA basis from registered 3D scans. Since 3D registrations are not available for in-the-wild videos, we optimize a parametric model through differentiable rendering. Given the stretched joint locations, the model deforms the canonical shape $\mathbf{T}_{\boldsymbol{\beta}}$ with blend skinning equations,

$$\mathbf{T}^s_{\boldsymbol{\beta}} = \big(\mathbf{W}_{\boldsymbol{\beta}}\mathbf{G}_{\boldsymbol{\beta}}\big)\mathbf{T}_{\boldsymbol{\beta}}, \quad (5)$$

where $\mathbf{G}_{\boldsymbol{\beta}}$ transforms the bone coordinates, and $\mathbf{W}_{\boldsymbol{\beta}}$ is the instance-specific skinning weights in Eq. (4).

3.2. Within-Instance Variation

We represent within-instance variations as time-varying warp fields between the canonical space and posed space at time t. Similar to HumanNeRF [57], we decompose motion as *articulations* that explains the near-rigid component (e.g., skeletal motion) and *deformation* that explains the remaining nonrigid movements (e.g., cloth deformation). Note given the complexity of body movements, it is almost certain the pre-defined skeleton would ignore certain movable body parts. Adding deformation is crucial to achieving high-fidelity reconstruction.

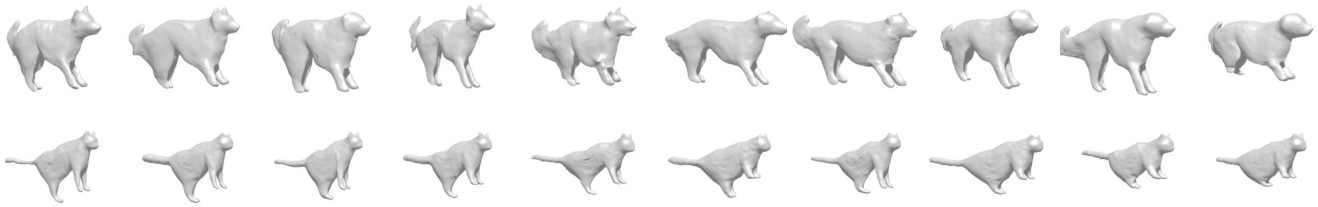

Figure 4. **Different $\boldsymbol{\beta}$ morphologies of dogs (top) and cats (bottom)**. Our reconstructions show variance in ear shape and limb size over dog breeds, as well as variance in limb and body size over cat breeds.

Time-varying Articulation. To model the time-varying skeletal movements, we define per-frame joint angles:

$$\mathbf{Q} = \mathbf{MLP}_{\mathbf{A}}(\boldsymbol{\theta}) \in \mathbb{R}^{3\times B} \tag{6}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{16}$ is a low-dimensional articulation parameter, and each joint has three degrees of freedom. Given joint angles and the per-video joint locations, we compute bone transformations $\mathbf{G} \in \mathbb{R}^{3\times 4\times B}$ via forward kinematics. We apply dual quaternion blend skinning (DQB) [19] to get the warping field for each spatial point,

$$\mathbf{D}(\boldsymbol{\beta}, \theta) = (\mathbf{W}_{\boldsymbol{\beta}}\mathbf{G})\mathbf{T}^{s}_{\boldsymbol{\beta}}. \tag{7}$$

Dual quaternion skinning blends SE(3) transformations in dual quaternion space and ensures valid SE(3) after blending, which reduces artifacts around twisted body parts. Note stretching in Eq. (5) can be fused with articulation as a single blend skinning operation.

Time-varying Soft Deformation. To further explain the dynamics induced by non-skeleton movements (such as the cat belly and human clothes), we add a neural deformation field [24, 40] $\mathcal{D}(\cdot)$ that is flexible enough to model highly nonrigid deformation. Applying the fine-grained warping after blend skinning, we have

$$\mathbf{D}(\boldsymbol{\beta}, \theta, \omega_d) = \mathcal{D}(\mathbf{D}(\boldsymbol{\beta}, \theta), \omega_d), \tag{8}$$

where ω_d is a frame-specific deformation code. Inspired by CaDeX [24], we use real-NVP [9] to ensure the 3D deformation fields are invertible by construction.

Invertibility of 3D Warping Fields. For a given time instance t, we have defined a forward warping field $\mathcal{W}^{t,\rightarrow}$ that transforms 3D points from the canonical space to the specified time instance, and a backward warping field $\mathcal{W}^{t,\leftarrow}$ to transform points in the inverse direction. Both warping fields include stretching (Eq. (5)), articulation. (Eq. (7)), and deformation (Eq. (8)) operations. Notably, we only need to define each operation in the forward warping fields. The deformation operation is, by construction, invertible. To invert stretching and articulation, we invert SE (3) transformations $\mathbf{G}$ in the blend skinning equations and compute the skinning weights with Eq. (4) using the corresponding morphology and articulation codes. A 3D cycle loss is used to ensure that the warping fields are self-consistent after a forward-backward cycle [29, 69].

3.3. Scene Model

Reconstructing objects from in-the-wild video footage is challenging due to failures in segmentation, which is often caused by out-of-frame body parts, occlusions, and challenging lighting. Inspired by background subtraction [15, 50], we build a model of the background to *robustify* our method against inaccurate object segmentation.

In background subtraction, moving objects can be segmented by comparing input images to a background model (e.g., a median image). We generalize this idea to model the background scene in 3D as a per-video NeRF, which can be rendered as color pixels at a moving camera frame and compared to the input frame. We design a conditional background model that generates density and color of a scene conditioned on a per-video background code $\boldsymbol{\gamma}$:

$$(\sigma, \mathbf{c}^t) = \mathbf{MLP}_{\mathbf{bg}}(\mathbf{X}, \mathbf{v}, \boldsymbol{\gamma}), \tag{9}$$

where $\mathbf{v}$ is the viewing direction. To render images, we compose the density and color of the object field and the background NeRF in the view space [37], and compute the expected color and optical flow. Background modeling and composite rendering allows us to remove the object silhouette loss, and improves the quality of results. Interestingly, we find that even *coarse* geometric reconstructions of the background still can improve the rendered 2D object silhouette, which in turn is useful for improving the quality of object reconstructions (Fig. 5). We ablate the design choice in Tab. 1.

3.4. Losses and Regularization

Given the videos and a predefined skeleton, we optimize the parameters discussed above: (1) canonical parameters $\{\boldsymbol{\beta}, \mathbf{J}, \mathbf{T}, \mathbf{W}\}$ including per-video morphology codes and canonical templates; (2) motion parameters $\{\boldsymbol{\theta}, \boldsymbol{\omega}_{\mathbf{d}}, \mathbf{A}, \mathbf{D}\}$ including per-frame codes as well as articulation and soft deformation MLPs. (3) background parameters $\{\boldsymbol{\gamma}, \mathbf{B}\}$ including video background codes and a background NeRF. The overall objective function contains a image reconstruction loss term and regularization terms.

Reconstruction Losses. The reconstruction loss is defined as the difference between rendered and observed images,

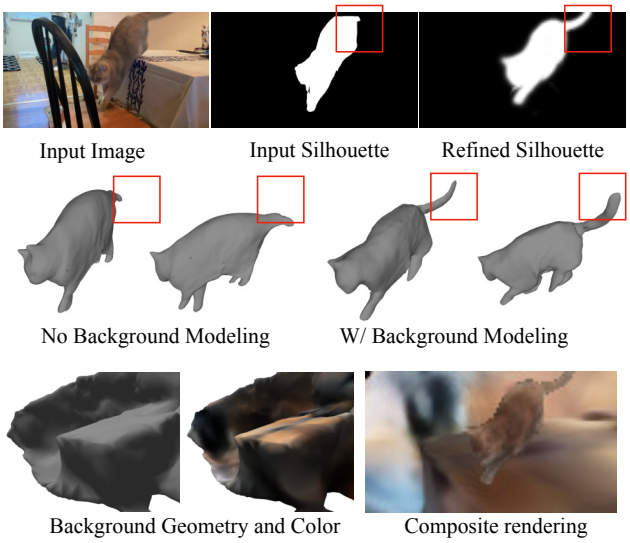

Figure 5. **Joint foreground and background reconstruction.** We jointly reconstruct objects and their background, while refining the segmentation. Note the input silhouette is noisy (e.g., tail was not segmented), and background modeling helps produce an accurate refined silhouette. As a result, RAC is robust to inaccurate segmentation (e.g., tail movements marked by the red box).

including object silhouette, color, flow, and features:

$$\mathcal{L} = \mathcal{L}_{\text{sil}} + \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{OF}} + \mathcal{L}_{\text{feat}}. \tag{10}$$

We update the model parameter by minimizing $\mathcal{L}$ through differentiable volume rendering in the same way as BANMo [69]. Off-the-shelf estimates of object silhouettes are used as supervision to kick-start the optimization. Then the weight of silhouette term is set to 0 after several iterations of optimization, while composite rendering of foreground and background itself is capable of separating the object and the non-object components.

Morphology Code ($\boldsymbol{\beta}$) Regularization. Existing differentiable rendering methods are able to faithfully reconstruct the input view but not able to hallucinate a reasonable solution for occluded body parts [40, 69]. See Fig. 9 for an example. One solution is to regularize the instance-specific morphology code $\boldsymbol{\beta}$ to be consistent with the body shapes observed in *other* videos. Traditional approaches might do this by adding variational noise (as in a VAE) or adversarial losses (as in a GAN). We found good results with the following easy-to-implement approach: we randomly *swap* the morphology code $\boldsymbol{\beta}$ of two videos during optimization; this regularizes the model to effectively learn a single morphology code that works well across all videos. But naively applying this approach would produce a single morphology that would not specialize to each object instance. To enable specialization, we *gradually* decrease the probability of swaps during the optimization, from $\mathcal{P} = 1.0 \rightarrow 0.05$.

Table 1. **Quantitative results on AMA sequences.** 3D Chamfer distance (cm, ↓) and F-score (%, ↑) averaged over all frames. Our model is trained on 47 human videos spanning existing human datasets (as described in Sec.4.2); we also train BANMo on the same set. Other baselines are trained on 3D human data and relies on SMPL model. Results with ^{S} indicates variants trained on single instances. Our model outperforms prior works.

Method	samba			bouncing		
	CD	F@2%	F@5%	CD	F@2%	F@5
HuMoR	9.8	47.5	83.7	11.5	45.2	82.3
ICON	10.1	39.9	85.2	9.7	53.5	86.4
BANMo^S	8.0	62.2	89.1	7.6	64.7	91.1
BANMo	9.3	54.4	85.5	10.2	54.4	86.5
RAC^S	6.4	70.9	93.2	**6.9**	**66.7**	**92.8**
RAC	**6.0**	**72.5**	**94.4**	8.0	63.8	91.4
w/o skeleton	8.6	59.6	87.7	9.3	59.5	87.8
w/o $\boldsymbol{\beta}$	8.5	58.9	87.5	8.4	62.5	90.6
$\boldsymbol{\beta}$ swap$\rightarrow$ $\|\|\boldsymbol{\beta}\|\|_2^2$	6.5	69.0	93.8	8.0	64.8	91.3
+ bkgd NeRF	6.3	70.9	93.7	7.4	65.5	91.8

Joint J Regularization. Due to the non-observable nature of the the joint locations, there might exist multiple joint configurations leading to similar reconstruction error. To register the skeleton with the canonical shape, we minimize Sinkhorn divergence [11] between the canonical surface $\mathbf{T}_{\boldsymbol{\beta}}$ and the joint locations $\mathbf{J}_{\boldsymbol{\beta}}$, which forces them to occupy the same space. We extract the canonical mesh with marching cubes [31] as a proxy of the canonical surface. Sinkhorn distance interpolates between Wasserstein and kernel distances and defines a soft way to measure the distance between shapes with different density distributions.

Soft Deformation Regularization The soft deformation field has the capacity of explaining not only the soft deformations, but also the skeleton articulations. Therefore, we penalize the L2 norm of the soft deformation vectors at randomly sampled morphology and articulations,

$$\mathcal{L}_{\text{soft}} = \|\mathbf{D}(\boldsymbol{\beta}, \theta, \omega_d) - \mathbf{D}(\boldsymbol{\beta}, \theta)\|. \tag{11}$$

4. Experiments

Implementation Details. We build RAC on BANMo and compute bone transformations from a kinematic tree. The soft deformation field follows CaDeX, where we find that two invertible blocks are capable of handling moderate deformations. To evaluate surface reconstruction accuracy, we extract the canonical mesh $\mathbf{T}$ by finding the zero-level set of SDF with marching cubes on a 256^3 grid. To get the shape at a specific time, the canonical mesh is forward-warped with $\mathcal{W}^{t,\rightarrow}$.

Optimization Details We use AdamW to optimize the model for 36k iterations with 16384 rays per batch (taking

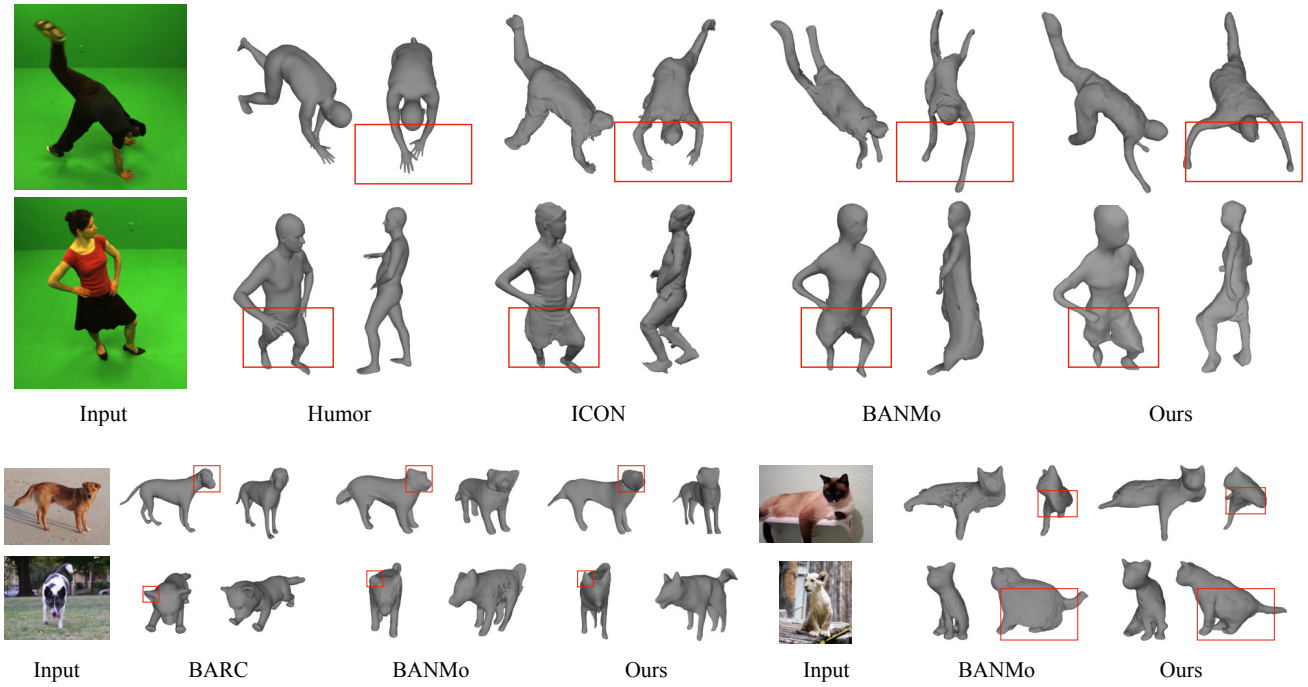

Figure 6. **Qualitative comparison.** We compare with BANMo and model-based methods (HuMoR, ICON, BARC). **Top**: human reconstruction on (AMA). **Bottom:** dogs and cats reconstruction on internet videos.

around 24 hours on 8 RTX-3090 GPUs). We fist pre-train the background model with RGB, optical flow, and surface normal losses while ignoring foreground pixels. Then we combine background models with the object model for composite rendering optimization. The weights for the loss terms are tuned to have similar initial magnitude. The object root poses are initialized with single-image viewpoint networks trained for humans and quadruped animals following BANMo [69]. For all categories, we start with the same shape (a unit sphere) and a known skeleton topology. Both the shape and the joint locations are specialized to the input dataset, as shown in Fig. 7.

4.1. Reconstructing Humans

Dataset. We combine existing human datasets, including AMA, MonoPerfCap, DAVIS, and BANMo to get 47 human videos with 6,382 images [43, 54, 64]. AMA contains multi-view videos, but we treat them as monocular videos and *do not* use the time-synchronization or camera extrinsics. During preprocessing, we use PointRend [20] to extract object segmentation, CSE [36] for pixel features, VCN-robust [66] for optical flow, and omnidata [10] for surface normal estimation.

Metrics. We use AMA for evaluation since it contains ground-truth meshes and follow BANMo to compute both Chamfer distances and F-scores. Chamfer distance computes the average distance between the ground-truth and the estimated surface points by finding the nearest-neighbor

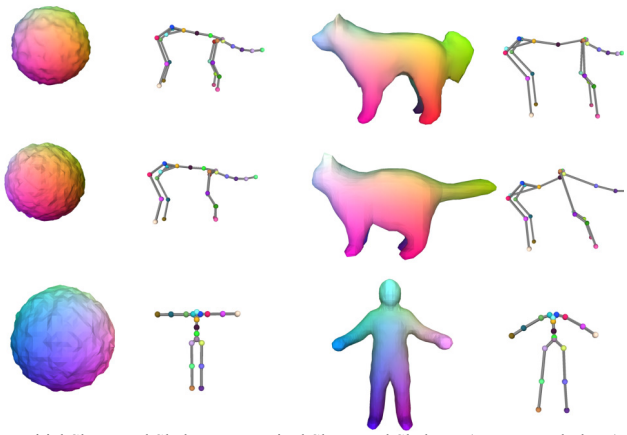

Figure 7. **Shape and skeleton optimization.** From top to bottom, we visualize the canonical shape and skeleton of our dog, cat, and human models. **Left**: Canonical shape and skeleton before optimization. **Right**: Canonical shape and skeleton after optimization.

matches, but it is sensitive to outliers. F-score at distance thresholds $d \in \{1\%, 2\%, 5\%\}$ of the human bounding box size [52] provides a more informative quantification of surface reconstruction error at different granularity. To account for the unknown scale, we align the predicted mesh with the ground-truth mesh using their depth in the view space.

Baselines. On AMA, we compare with template-free BANMo [69] and model-based methods, including Hu-

MoR [45] and ICON [63]. BANMo reconstructs an animatable 3D model from multiple monocular videos of the same instance, powered by differentiable rendering optimization. We optimize BANMo on the same dataset with the same amount of computation and iterations as ours. HuMoR is a temporal pose and shape predictor for humans. It performs test-time optimization on video sequences leveraging OpenPose keypoint detection and motion priors trained on large-scale human motion capture dataset. We run it on each video sequence, and processing 170 video frames takes around two hours on a machine with Titan-X GPU. ICON is the recent SOTA method on single view human reconstruction. It combines statistical human body models (SMPL) with implicit functions and is trained on 3D human scans with clothing. Notably, it performs test-time optimization to fit surface normal predictions to improve the pose accuracy and reconstruction fidelity. We run it per frame, and processing a 170 frame video takes around three hours on an RTX-3090 GPU.

Results. We show qualitative comparison with baselines in Fig. 6 top row, and quantitative results in Tab. 1. On the handstand sequence, HuMoR works well for common poses but fails where the performer is not in an upright pose. ICON works generally well, but the hand distances appear not physically plausible (too short) from a novel viewpoint. BANMo reconstruction also failed to reconstruct the unnatural upside-down pose. In contrast, RAC successfully reconstructs the handstand pose with plausible hand distances. On the samba sequence, HuMoR correctly predicts body poses, but fails to reconstruct the cloth and its deformation. ICON predicts a broken dress and distorted human looks from a novel viewpoint, possibly due to lack of diverse training data from dressed humans. When applied to 47 videos of different humans, BANMo fails to model the cloth correctly, possibly because a limited number of control points are not expressive enough to model the morphological variations over humans wearing different clothes. RAC models between-shape variations using a conditional canonical model and successfully reconstructs cloth deformation using the soft deformation field.

Our quantitative results align with qualitative observations, where RAC outperforms all baselines except being slightly worse than BANMo trained on single instances (S). However, RAC trained on single instances (S) or multiple instances (M) outperforms BANMo trained in either fashion. In particular, BANMo results are notably worse when trained on multiple instances, indicating the difficulty in building category models. In contrast, RAC become better when trained on multiple instances.

4.2. Reconstructing Cats and Dogs

Dataset. We collect 76 cat videos and 85 dog videos from Internet videos, as well as public data from BANMo. All the videos are casually-captured monocular videos. We extract video frames at 10 FPS, including 9,734 frames for cats and 11,657 frames for dogs. We perform the same preprocessing as human reconstruction.

Baselines. We compare with BANMo and model-based BARC [46]. BARC is the current SOTA for dog shape and pose estimation. It trains a feed-forward network using CG dog models and images with keypoint labels. The shape model is based on SMAL, which uses manual rigging and registration to fit 3D animal toys. We run BARC on individual images.

Results. We show qualitative results in Fig. 6 bottom row. For dog videos, we find that BARC worked well to predict coarse shapes and poses. However, the results are biased towards certain dog breeds. For instance, BARC predicts a long jaw when the dog has a short jaw (top row), and predicts round ears when the dog has sharp ears (bottom). BANMo was able to reconstruct a reasonable coarse shape, but failed to capture the fine details (e.g., the shape of the ear and the size of the head) with only 25 control points. In contrast, RAC was able to faithfully capture the coarse shape and fine details of the dogs. Unlike dogs, cats have fewer variations in body shape and size, where we find that BANMo works well in most cases. However, for the body parts not visible from the reference viewpoint, BANMo often estimates a squashed shape, which may be caused by the entangled morphology and articulations. In contrast, RAC accurately infers reasonable body parts and articulations even when they are not visible.

4.3. Diagnostics

Large Morphological Changes. We reconstruct eight videos of different quadruped animals together to "stress test" our method. The dataset contains two dog videos, two cat videos, and one video for goat, bear, camel, and cow, respectively. The result is shown in Fig. 8.

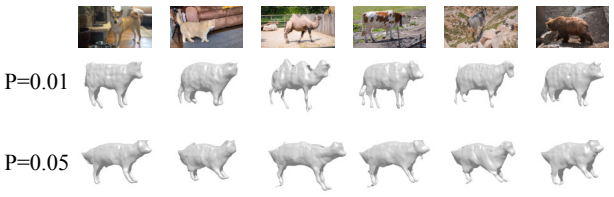

Figure 8. **Quadruped Category Reconstruction.** Using a smaller code swapping probability $P = 0.01$ results in more faithful instance shape, but less smooth results. A larger P produces smoother results, but some instance-specific features disappear.

Morphology Code β Removing morphology code β from the canonical field degrades it to a standard NeRF. We rerun the experiments and the results are shown in Tab. 1 as well as Fig. 9. Without the morphology code, our reconstructions are forced to share the same canonical shape, which

as discussed in Sec 3.1, failed to handle fine deformations and topological changes (e.g., clothing), leading to worse results in all metrics.

Morphology Code Regularization To test the effectiveness of the morphology code regularization, we set $\mathcal{P} = 0$ throughout the optimization. The results are shown in Tab. 1 and Fig. 9. Without regularization of the morphology code, the reconstructed shape may appear reasonable from the reference viewpoint, but severely distorted from a novel viewpoint. We posit the body parts that are not well-covered in the video are inherently difficult to infer. As a result, the shapes become degenerate without relying on priors from other videos in the dataset. Tab. 1 also shows that code swapping outperforms norm regularization [12] ($||\boldsymbol{\beta}||_2^2$). We posit norm regularization forces codes to be similar, but does not constrain their output space, while code swapping encourages *any* code to explain *any* image in the data. In practice, we find that code swapping generates valid outputs when we interpolate between codes.

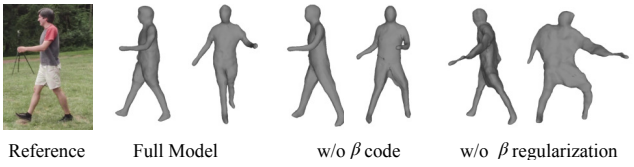

Figure 9. **Ablation study on morphology modeling**.

Soft Deformation. After removing the soft deformation field, RAC fails to recover body parts that are not controlled by the skeleton (Fig. 10), such as the ears of the dog.

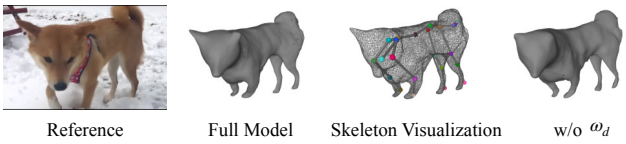

Figure 10. **Ablation study on soft deformation $\boldsymbol{\omega}_d$**.

Motion Transfer. Given the category model with disentangled morphology and articulations, we can easily transfer an articulation in frame t to other video by setting the articulation parameter to $\boldsymbol{\theta}_t$, while keeping the morphology parameter $\boldsymbol{\beta}$ the same. We show motion transfer across human in Fig. 2. Please visit our website for video results of dogs, cats, and humans.

Skeleton vs Control Points. Control point deformations are flexible but do not preserve body dimensions (e.g., a line segment can be stretched longer by its end points). As a result, body and limb dimensions can change, creating two problems: (1) articulated shapes look squashy from novel views, and (2) variations in body dimensions are entangled with control-point deformations. In contrast, skeleton deformation preserves body dimensions. It produces better results (Tab. 1) and better motion re-targeting (Fig. 11).

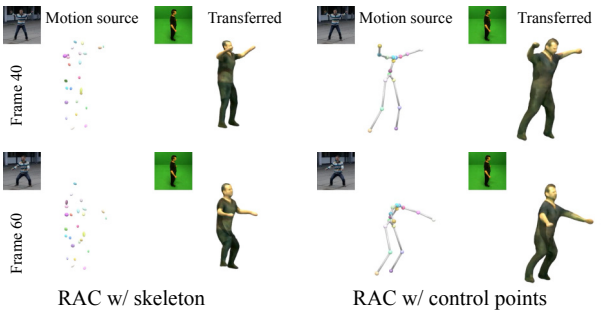

Figure 11. **Skeleton vs control-points for motion transfer**. RAC with control points fails to maintain the body dimensions during motion, and produces squashy results when transferred to a new morphology. RAC with skeleton disentangles motion from morphology, allowing for better motion transfer.

Stretchable Bones allow for control of bone dimensions after optimization. We show an example of a Dachshund (Source1) warped to a Heeler (Source2) by modifying bone dimensions while keeping the shape unchanged.

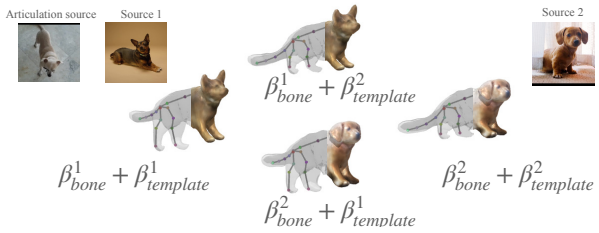

Figure 12. Disentanglement of bone dimensions and shape.

5. Discussion

We have presented a scalable way of building animatable category models by learning from monocular videos. It disentangles morphology variations between instances and motion within an instance, allowing motion transfer over a category. RAC reaches state-of-the-art reconstruction quality for cats, dogs, and humans in terms of mid-level reconstruction, but details are still missing (such as human hand and toes). Similar to BANMo, RAC requires rough viewpoint initialization. Although we have shown either a pretrained viewpoint estimator or roughly annotated camera viewpoints (in the supplement) are sufficient, it would be interesting to study a more generic way to initialize viewpoints. We also show that category-level skeleton improves motion reconstruction, and leave jointly inferring skeleton structure together with object shape for future work.

Acknowledgement This work was supported by the Qualcomm Innovation Followship and the CMU Argo AI Center for Autonomous Vehicle Research.

References

[1] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 1

[2] Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd Pfrommer, Marc Schmidt, and Kostas Daniilidis. 3D bird reconstruction: a dataset, model, and shape recovery from a single view. In *ECCV*, 2020. 2

[3] Benjamin Biggs, Ollie Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out: 3D animal reconstruction with expectation maximization in the loop. In *ECCV*, 2020. 2

[4] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and smal: Recovering the shape and motion of animals from video. In *ACCV*, pages 3–19. Springer, 2018. 1

[5] Federica Bogo, Javier Romero, Matthew Loper, and Michael J Black. Faust: Dataset and evaluation for 3d mesh registration. In *CVPR*, pages 3794–3801, 2014. 1

[6] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *arXiv*, 2020. 2

[7] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *arXiv*, 2021. 2

[8] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, pages 5799–5809, 2021. 3

[9] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 4

[10] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, pages 10786–10796, 2021. 6

[11] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019. 5

[12] Partha Ghosh, Medhi SM Sajjadi, Antonio Vergari, and Michael Black. From variational to deterministic autoencoders. In *ICLR*, 2020. 8

[13] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoints without keypoints. In *ECCV*, 2020. 2

[14] Alec Jacobson and Olga Sorkine. Stretchable and twistable bones for skeletal shape deformation. In *SIGGRAPH Asia*, pages 1–8, 2011. 3

[15] Ramesh Jain and H-H Nagel. On the analysis of accumulative difference pictures from image sequences of real world scenes. *TPAMI*, (2):206–214, 1979. 4

[16] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *ICCV*, pages 12949–12958, October 2021. 3

[17] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022. 2

[18] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2

[19] Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Skinning with dual quaternions. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, pages 39–46, 2007. 4

[20] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, 2020. 6

[21] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, June 2020. 2

[22] Filippos Kokkinos and Iasonas Kokkinos. To the point: Correspondence-driven monocular 3d category reconstruction. In *NeurIPS*, 2021. 2

[23] Binh Huy Le and Zhigang Deng. Robust and accurate skeletal rigging from mesh sequences. *ACM Transactions on Graphics (TOG)*, 33(4):1–10, 2014. 3

[24] Jiahui Lei and Kostas Daniilidis. Cadex: Learning canonical deformation coordinate space for dynamic surface representation via neural homeomorphism. In *CVPR*, 2022. 4

[25] Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li. Volumetric human teleportation. In *ACM SIGGRAPH 2020 Real-Time Live!*, pages 1–1. 2020. 2

[26] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. 2022. 2

[27] Xueting Li, Sifei Liu, Shalini De Mello, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Online adaptation for consistent mesh reconstruction in the wild. In *NeurIPS*, 2020. 2

[28] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. *ECCV*, 2020. 2

[29] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. 4

[30] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 1, 2, 3

[31] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH*, 21(4):163–169, 1987. 5

[32] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duck-

worth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 3

[33] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2

[34] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[35] Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. In *NeurIPS*, 2020. 3

[36] Natalia Neverova, Artsiom Sanakoyeu, Patrick Labatut, David Novotny, and Andrea Vedaldi. Discovering relationships between object categories via universal canonical maps. In *CVPR*, 2021. 6

[37] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, pages 11453–11464, 2021. 2, 3, 4

[38] Atsuhiro Noguchi, Umar Iqbal, Jonathan Tremblay, Tatsuya Harada, and Orazio Gallo. Watch it move: Unsupervised discovery of 3D joints for re-posing of articulated objects. *CVPR*, 2021. 3

[39] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, 2021. 2

[40] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021. 1, 4, 5

[41] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv*, 2021. 1, 3

[42] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 2

[43] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 6

[44] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *CVPR*, 2020. 1

[45] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, 2021. 7

[46] Nadine Rueegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. BARC: Learning to regress 3D dog shape from images by exploiting breed information. In *CVPR*, pages 3876–3884, 2022. 2, 7

[47] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, pages 2304–2314, 2019. 2

[48] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 2

[49] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, 2021. 2

[50] Yaser Sheikh, Omar Javed, and Takeo Kanade. Background subtraction for freely moving cameras. In *ICCV*, pages 1219–1225. IEEE, 2009. 4

[51] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, 2021. 2

[52] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, pages 3405–3414, 2019. 6

[53] Shubham Tulsiani, Nilesh Kulkarni, and Abhinav Gupta. Implicit mesh reconstruction from unannotated image collections. In *arXiv*, 2020. 2

[54] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *SIGGRAPH 2008*. 2008. 6

[55] Minh Vo, Yaser Sheikh, and Srinivasa G Narasimhan. Spatiotemporal bundle adjustment for dynamic 3d human reconstruction in the wild. *IEEE TPAMI*, 2020. 2

[56] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. In *arXiv preprint arXiv:2102.07064*, 2021. 1

[57] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, pages 16210–16220, June 2022. 3

[58] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Dove: Learning deformable 3d objects by watching videos. *arXiv preprint arXiv:2107.10844*, 2021. 2

[59] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. *arXiv preprint arXiv:2211.12497*, 2022. 2

[60] Yuefan Wu, Zeyuan Chen, Shaowei Liu Liu, Zhongzheng Ren, and Shenlong Wang. Casa: Category-agnostic skeletal animal reconstruction. In *NeurIPS*, 2022. 3

[61] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019. 2

[62] Christopher Xie, Keunhong Park, Ricardo Martin-Brualla, and Matthew Brown. Fig-nerf: Figure-ground neural radiance fields for 3d object category modelling. In *3DV*, pages 962–971. IEEE, 2021. 3

[63] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained

from Normals. In *CVPR*, pages 13296–13306, June 2022. 2, 7

[64] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Trans. Graph.*, 37(2):27:1–27:15, May 2018. 6

[65] Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Landreth, and Karan Singh. Rignet: Neural rigging for articulated characters. *ACM Trans. on Graphics*, 39, 2020. 3

[66] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. In *NeurIPS*, 2019. 6

[67] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. LASR: Learning articulated shape reconstruction from a monocular video. In *CVPR*, 2021. 2

[68] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. In *NeurIPS*, 2021. 2, 3

[69] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 1, 2, 4, 5, 6

[70] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *CVPR*, 2021. 2

[71] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2

[72] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *ICCV*, 2019. 2

[73] Silvia Zuffi, Angjoo Kanazawa, and Michael J. Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *CVPR*, 2018. 2

[74] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*, 2017. 2, 3