Animal Avatars: Reconstructing Animatable 3D Animals from Casual Videos

Remy Sabathier[1,2], Niloy J. Mitra[2], and David Novotny[1]

[1] Meta, London, United-Kingdom
{rsabathier,dnovotny}@meta.com
[2] University College London (UCL), London, United-Kingdom
n.mitra@ucl.ac.uk

Abstract. We present a method to build animatable dog avatars from monocular videos. This is challenging as animals display a range of (unpredictable) non-rigid movements and have a variety of appearance details (e.g., fur, spots, tails). We develop an approach that links the video frames via a 4D solution that jointly solves for animal's pose variation, and its appearance (in a canonical pose). To this end, we significantly improve the quality of template-based shape fitting by endowing the SMAL parametric model with Continuous Surface Embeddings (CSE), which brings image-to-mesh reprojection constaints that are denser, and thus stronger, than the previously used sparse semantic keypoint correspondences. To model appearance, we propose a novel implicit duplex-mesh texture that is defined in the canonical pose, but can be deformed using SMAL pose coefficients and later rendered to enforce a photometric compatibility with the input video frames. On the challenging CoP3D and APTv2 datasets, we demonstrate superior results (both in terms of pose estimates and predicted appearance) over existing template-free (RAC) and template-based approaches (BARC, BITE). Video results and additional information accessible on the project page: https://remysabathier.github.io/animalavatar.github.io.

1 Introduction

Building poseable reconstructions of animals captured by consumer imaging devices is a valuable technology with numerous applications in augmented and virtual reality. Among many possible animal species, the reconstruction of canines is of particular interest primarily due to their important role in the lives of their two-legged friends.

While it is nowadays possible to reliably reconstruct rigid scenes captured from a moving vantage point [41], the reconstruction of non-rigid shapes is significantly less constrained and, as such, a more challenging problem. Here, many recent works focused on generic animal reconstruction from multi-view (videos) [19,52,54] or single-view 2D image supervision [49] without prior knowledge of the animal shape. While the latter demonstrated impressive progress, the challenging nature of the problem limits the applicability to scenarios with simple deformations and good multi-view test-time coverage.

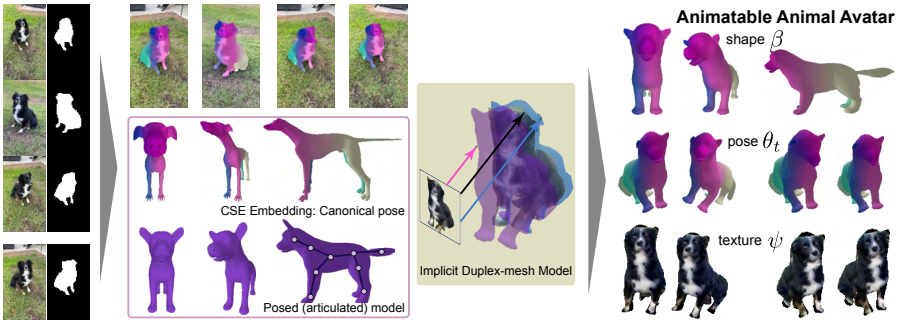

Fig. 1: Animal Avatars. Given a monocular video of a dog, we propose a template-based method to reconstruct the shape β, motion θ_t and texture ψ. We address the challenge of insufficient supervision for unconventional views by integrating Continuous Surface Embeddings with an articulated mesh. We introduce a novel implicit duplex-mesh texture model, jointly optimized alongside motion parameters.

To better constrain the non-rigid 3D reconstruction, inspired by state-of-the-art methods for Human 3D reconstruction [7, 27], we leverage a known category template. Specifically, we use the SMAL template introduced by Zuffi *et al.* [62] - a quadruped-equivalent of the seminal SMPL parametric human model [27]. The template has been used to enable single-view 3D dog reconstruction trained on an extensive collection of dog images semi-manually annotated with SMAL poses [8, 39, 40]. While this approach provides a clear state-of-the-art in monocular 3D dog-shape reconstruction, the inherent ambiguity of single-view reconstruction provides many challenges.

To further increase our chances of successful reconstruction, besides leveraging the SMAL model, we turn our attention towards reconstructing casual video captures of dogs, which was first explored in [3]. Indeed, videos provide a stronger multi-view supervision which significantly simplifies the 3D shape fitting problem. However, regardless of monocular [39, 40] or multi-view conditioning [3], we observed a common failure mode in existing methods when animals are viewed from non-frontal views. This issue arises because fitting relies on sparse joint-reprojection constraints that mainly cover front-facing parts, offering limited supervision for rear and side views.

Hence, our first contribution is to replace the sparse keypoint supervision with a denser alternative. Specifically, we exploit Continuous Surface Embeddings [31] (CSE), which annotate *each* vertex of the CSE dog mesh with a unique descriptor. Furthermore, CSE provides a pre-trained deep image-to-CSE predictor, that labels image pixels with their corresponding CSE descriptor and transitively with the matching mesh vertex. In this work, in an one-time process, we first transform CSE descriptors to the SMAL mesh by means of a semi-manual non-rigid aligment. This then enables a stronger keypoint loss providing reprojection constraints for all points on the animal's body, even in rear views.

Secondly, we enhance fits by exploiting the inherent smoothness of animal movements over time. Previous attempts incorporated this knowledge by enforcing temporal smoothness on the deformation coefficients of the SMAL template. However, we found this approach flawed because the coefficients have to represent both the smooth animal motion and the camera motion, which is often non-smooth due to the unstable camera operator. Instead, we propose to represent the SMAL deformation as a combination of accurate Structure-from-Motion camera and optimized animal motion, allowing for proper temporal regularization. Importantly, SfM also provides intrinsic parameters of each camera (focal length), which facilitates more accurate shape fitting through loss terms that require rendering.

Finally, we are the first to enable texturing of the SMAL mesh by leveraging it as a scaffold for a novel implicit duplex-mesh neural radiance field, which can be rendered while accounting for body deformations. Our approach involves defining implicit shape and color functions on a subset of the 3D domain bounded by enlarged and downsized versions of the mesh template. This allows for articulation of the corresponding implicit surface by posing the boundary meshes similarly to the original mesh.

We evaluate the color and 3D shape-fitting accuracy on the CoP3D dataset [42], containing crowd-sourced "turntable" videos of dogs captured by smartphone cameras, achieving performance superior to template-based [39, 40] and template-free [55] baselines. We also compare our pose estimation quality on the recent APTv2 dataset [56] and report results superior compared to video-based methods dog-specific [55].

2 Related Work

Video reconstruction on humans. Recent works in 3D human pose reconstruction show impressive results on videos, representing detailed motions and robustness to occlusions [10,59]. Several factors support the progress in this area.

The majority of methods rely on the parametric *SMPL* model introduced in [27] and refined in [33,34,38]. Unlike the existing animal models, it is learned from a large collection of real 3D scans of humans, which entails stronger expressiveness. The model provides parametric handles to both shape and pose variations. Additionally, human-centric models benefit from large real datasets with keypoint annotations [24, 28, 48] as well as synthetic 3D datasets [13].

We also note the availability of off-the-shelf models for related tasks such as human key-point identification [51]. Such models can guide training of 3D reconstruction or provide soft annotation on large unlabelled datasets [11]. These factors explain why 3D animal reconstruction cannot directly benefit from breakthroughs in the human domain.

Additionally, there are several works [6, 12, 15, 35] targeting human reconstruction with texture from monocular and multi-view sources achieving high rendering quality by leveraging off-the-shelf human pose estimation models.

Template-based animal reconstruction. Based on the success of template-based human reconstruction using SMPL, [62] introduced *SMAL*, a parametric model for quadruped animals. Unlike the SMPL model, which is supervised by scans of real humans, and due to the many challenges of scanning live quadruped animals, the SMAL model is only trained with scans of toy animal models.

To add to the challenge, datasets with 3D annotations for dog reconstruction are very limited [17, 50] and does not adequately represent the diversity of dog breeds and poses. The most relevant image dataset is *StanfordExtra* [8, 18], a collection of dog images with silhouette and joint annotations. Despite being diverse, spanning different dog breeds and environments, the dataset is biased towards front-facing views. This motivates our choice of the CoP3D dataset [42], an extensive collection of pet videos with high view-point variability in each video.

Similar to human reconstruction research, multiple works estimate shape attributes β and pose attributes θ for the SMAL model from a *single* image, relying on 2D reprojection constraints. [3] predicts skeleton joints to find an initial solution, which is then refined to match keypoints estimated from the animal silhouette. [22] propose a coarse-to-fine strategy, where an initial solution is refined through a graph-convolutional network. BARC [39] enforces similar shape attributes for dogs of the same *breed*, which is predicted by a deep network. BITE [40], an extension of BARC, improve prediction plausibility with *ground-contact* and *ground-plane* losses, and improve accuracy with an iterative refinement loop. We compare ours against BARC and BITE to show how multi-view supervision and CSE embeddings significantly improve performance, especially on the challenging videos from the COP3D dataset. We note some additional template-based related works [1, 2, 46, 57, 61].

Template-free animal reconstruction. The template-free setup enabling reconstruction of a wider range of animals has also been considered. These approaches build a 3D representation by analyzing a collection of images, or frames of a single video, or videos of the same species.

Progress made in differentiable rendering [25, 37, 47] supported the *analysis-by-synthesis* method for animal reconstruction from a single video. Recent works [19, 32, 52, 53] propose to minimize silhouette and photometric losses in order to jointly learn camera parameters and a textured frame-independent linear deformation 3D model.

We are inspired by Viser [53] that recovered articulated humans from monocular videos. They learn long-range 2D point tracks using object masks and optical flow, and a video-specific embedding linking pixel appearance to surface points on a canonical deformable mesh. However, such approaches are vulnerable to (self-)occlusions, especially in significantly dynamic scenes.

To overcome these issues, several works train on a set of videos. *BANMo* [54] leverages multiple videos of the same subject to build an implicit canonical representation, posed via a differentiable *neural blend skinning* method. Similar to our work, they use pixel *CSE* predictions on the different images to link it with a 3D embedding defined in the canonical space. As such, satisfactory

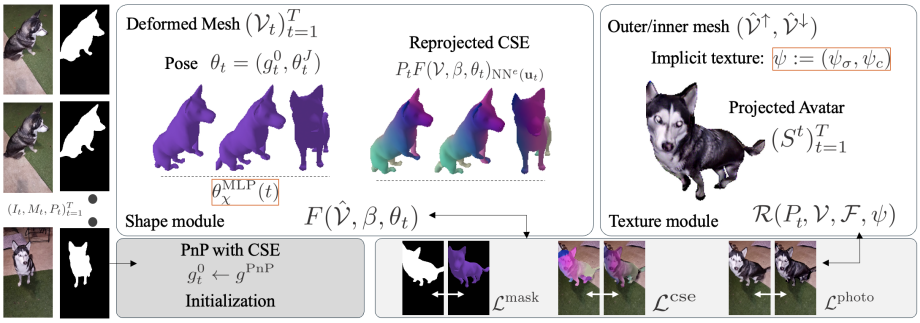

Fig. 2: Method overview. Two stage process: *First*, we initialize root pose g_t^0 via PnP-RANSAC, utilizing CSE mesh-pixel correspondences. *Then*, we jointly optimize shape β, time-varying pose θ_t and implicit texture ψ through an analysis-by-synthesis approach, leveraging mask $\mathcal{L}^{mask}$, dense correspondence $\mathcal{L}^{cse}$ (optimized models in orange), and photometric $\mathcal{L}^{photo}$ signals.

new views are generated only when positioned relatively close to the training views. Recently, RAC [55] extended BANMo to learn a general and instance-specific model from a set of videos from the same category of deformable objects, including the dog category analyzed here. TrackeRF [42] extends PixelNeRF [58] to time-deforming shapes but it only predicts novel views without a full-scale 3D animal-body model. Note that the method additionally needs to be initialized with a basic hierarchical joint skeleton. We compare with RAC in Sec. 4 and omit comparison to TrackeRF because its source code is not available.

Also note [14, 23] both propose innovative methods for reconstructing 3D models from 2D data, with the former developing a deformable 3D model for various species using unlabelled images, and the latter creating category-specific reconstructors through supervision from a diffusion image generator.

3 Method

Our goal is to reconstruct the 3D shape and appearance of a dog captured in a monocular video, i.e., given a tuple $(I_t)_{t=1}^T$ of $T \in \mathbb{N}$ video frames we output a tuple $(S_t)_{t=1}^T$ of colored 3D shapes S_t of the animal in each frame $I_t \in \mathbb{R}^{3\times H\times W}$.

In Sec. 3.1, we detail the shape representation S, while Secs. 3.2 and 3.3 describe the optimization process recovering the shape S given input images I.

3.1 Shape and Appearance Representation

Our method defines the colored shape representation S_t at time t as a 3-tuple

$$S_t := (\beta, \theta_t, \psi), \tag{1}$$

where β and ψ define the intrinsic (i.e., time-invariant) deformation and the texture of the dog body respectively, and θ_t is the time-dependent pose of its skeleton. In what follows, we detail these three sets of parameters.

3D shape representation β, θ_t. There is a plethora of articulated 3D shape representations ranging from universal less-constrained 3D-flow functions [9,36] to category-specific Linear-Blend-Skinning (LBS) models attached to a fixed surface-mesh template [21,27,62]. Since we focus on a certain animal category (i.e., dog) with a well-defined space of plausible articulations and body deformations, we opt for the latter, i.e., a template-based shape model.

Specifically, we represent the 3D geometry (i.e., the parameters θ_t, β in Eq. (1)) of each dog with the SMAL model [62]. The latter comprises a deformable template mesh $(\hat{\mathcal{V}}, \mathcal{F})$ with vertices $\hat{\mathcal{V}} \in \mathbb{R}^{3889\times 3}$ and a list of triangular faces $\mathcal{F} \in \mathbb{N}^{7774\times 3}$. The deformation of $\hat{\mathcal{V}}$ is defined as a function

$$F(\hat{\mathcal{V}}, \beta, \theta_t) := \mathcal{V} \in \mathbb{R}^{3889\times 3}, \tag{2}$$

conditioned on shape parameters (PCA coefficients and bone lengths) $\beta \in \mathbb{R}^{d_\beta}$, controlling the non-rigid deformation of the animal shape in its *canonical* pose, and pose parameters $\theta_t := (g^0, \theta^J)$. The latter has two components: (i) a root transformation $g^0 \in \mathbb{SE}(3)$ that represents the overall rigid transformation of the dog body; and (ii) angles $\theta^J \in \mathbb{R}^{d_J}$ of the animal joints that control the deformation of its limbs.

Time-deforming SMAL. To represent the time-varying deformation of a dog, we estimate a tuple $(\theta_t)_{t=1}^T$ comprising SMAL pose coefficients θ_t for each of the T video frames, and a single vector β because, typically, the intrinsic deformation is time-invariant. Since the animals often move smoothly in time, the pose θ_t is defined as a function $\theta^{\text{MLP}}(t)$ of a smooth temporal basis $\gamma(\tau(t))$ as follows:

$$\theta_t := \theta^{\text{MLP}}(\tau(t)), \tag{3}$$

implemented by a shallow multi-layer perceptron θ^{MLP} accepting positional encoding γ [44] applied to the timestamp $\tau(t) \in \mathbb{R}^+$ of frame I_t.

Continuous SMAL Embeddings. Besides the deformation parameters (θ, β), SMAL also defines joint locations $\hat{\mathcal{J}} \in \mathbb{R}^{N_J\times 3}$ as a sparse subset of 3D points (not necessarily localized on the surface of the parametric shape), which are linearly regressed from the vertex locations. Since the joints correspond to semantic parts of the animal body (paws, ear tips), they can improve shape-fitting accuracy via establishing correspondences between the SMAL mesh and the detections of the body parts in the images. However, while these keypoints improve performance in [39,40], where most animals are photographed from their side or frontal views, they are inadequate for our video fly-arounds containing rear views with little-to-no visible keypoints (see Fig. 3 and Sec. 4).

Hence, instead of sparse landmarks, we exploit Continuous Surface Embeddings (CSE) [31], which attach a unique embedding vector $e_k \in \mathbb{R}^{d_e}$ to *each* vertex $X_k \in \hat{\mathcal{V}}$ of the (canonical) SMAL template such that the dimensions of e vary smoothly over the mesh surface. CSE also provides a deep predictor that annotates each image pixel with its corresponding mesh coordinate e_k. Thus, unlike sparse keypoints, the latter densely annotates images of animals from arbitrary viewpoints, including rear views.

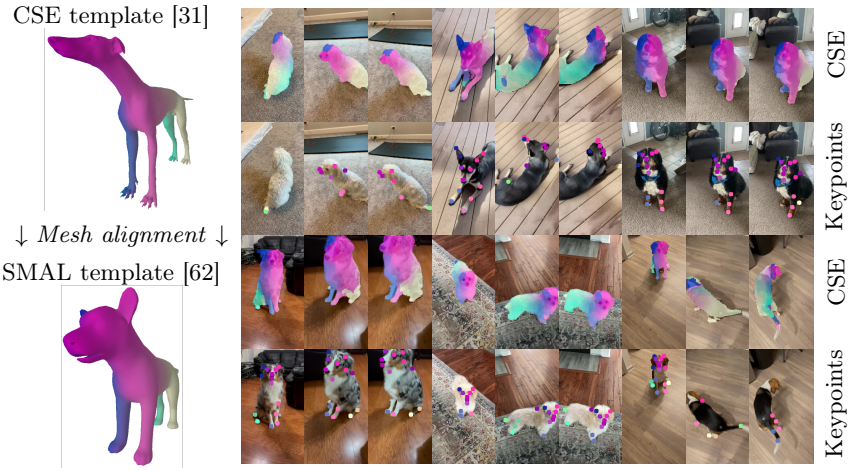

Fig. 3: SMAL CSE. We align the CSE mesh template (top left) with the SMAL template (top right) in order to setup the CSE coordinate frame over the surface of the SMAL mesh. In combination with a pretrained image-to-CSE predictor, this allows establishing dense correspondences between surface points of the SMAL template and the corresponding image pixels. Note that the image CSE detections (rows labelled "CSE") provide dense correspondence covering all parts of the dog body, which is not the case of sparse keypoints ("Keypoints" rows).

Because the original template mesh of CSE [31] is different from the SMAL template, following [31], we transform the CSE coordinate map to SMAL using a customized variant of the Zoom-Out method [29]. The latter results in the final set of SMAL-CSE vertex coordinates e_k (alignment visualized in Fig. 3).

Dual-mesh implicit texture ψ. Besides reconstructing 3D animal shapes, we also aim to recover the texture of the animal body. We require an exact supporting 3D shape to learn an accurate texture model. However, due to the low expressivity of the SMAL deformation space, the posed mesh can only represent the surface of a dog up to a certain error. Thus, following recent advances in new-view synthesis of humans [26], we leverage the template mesh as a scaffold supporting a more accurate implicit radiance field [30], which we describe next.

Our method is inspired by duplex radiance fields [45], originally proposed for speeding up rendering of neural radiance fields [30]. Specifically, we first extrude the 2-manifold canonical surface to a 3D volume by defining an ϵ-neighborhood $\hat{\mathcal{N}}^\epsilon \subset \mathbb{R}^3$ as a 3D subspace bounded by an *outer* mesh with vertices $\hat{\mathcal{V}}^\uparrow = (1+\epsilon)\hat{\mathcal{V}}$, and an *inner* mesh $\hat{\mathcal{V}}^\downarrow = (1-\epsilon)\hat{\mathcal{V}}$, both sharing the same faces $\mathcal{F}$. In practice, we obtain the offset meshes by moving along directions of vertex normals. Similar to the template mesh itself, both $\hat{\mathcal{V}}^\uparrow$ and $\hat{\mathcal{V}}^\downarrow$ can be deformed with θ, β resulting in the posed neighborhood $\mathcal{N}^\epsilon$ bounded by $F(\hat{\mathcal{V}}^\uparrow, \beta, \theta)$ and $F(\hat{\mathcal{V}}^\downarrow, \beta, \theta)$.

Within this neighborhood, we then define an opacity function $\psi_\sigma : \mathcal{N}^\epsilon \mapsto \mathbb{R}^+$, annotating 3D locations $\hat{\mathbf{x}}$ in the mid-space with presence ($\psi_\sigma(\hat{\mathbf{x}}) > 0$) or absence ($\psi_\sigma(\hat{\mathbf{x}}) \approx 0$) of the surface, and the radiance function $\psi_c : \mathcal{N}^\epsilon \times \mathbb{S}^2 \mapsto [0,1]^3$,

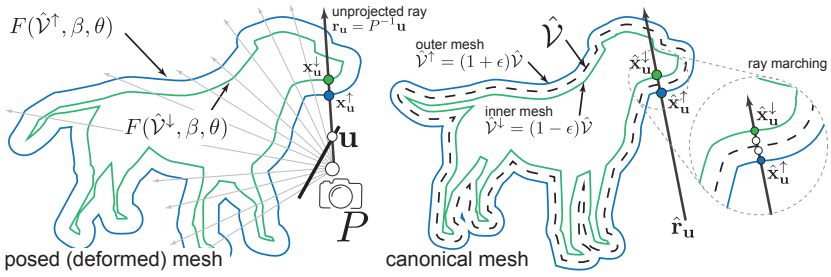

Fig. 4: Implicit duplex-mesh model. We propose a novel deformable implicit shape model. Using the radiance and opacity functions ψ_c and ψ_σ defined over an $\mathbb{R}^3$ subspace bounded by a canonical duplex mesh with vertices $\hat{\mathcal{V}}^{\uparrow}$ and $\hat{\mathcal{V}}^{\downarrow}$, we render a color of the posed duplex mesh via EA raymarching over a canononical ray $\hat{\mathbf{r}}_\mathbf{u}$. The latter is obtained by transforming into the canonical space the intersections of the view-space ray $\mathbf{r}_\mathbf{u}$ with the posed duplex mesh $F(\hat{\mathcal{V}}^{\uparrow}, \beta, \theta), F(\hat{\mathcal{V}}^{\downarrow}, \beta, \theta)$.

which colors the space depending on the direction $\mathbf{r} \in \mathbb{S}^2$ from which the input point $\hat{\mathbf{x}}$ is observed. The coloring and opacity functions are implemented as in [5], i.e., using a shallow MLP decoding a learnable triplane feature grid. Please refer to the supplementary material for details.

Dual-mesh rendering. Having defined the animal shape $(\mathcal{V}, \mathcal{F})$ and the implicit texture ψ, we image the latter from an arbitrary camera viewpoint P using a differentiable rendering function $\mathcal{R}$:

$$\mathcal{R}(P, \mathcal{V}, \mathcal{F}, \psi) := \bar{I}, \tag{4}$$

which outputs the render $\bar{I} \in [0,1]^{3\times H\times W}$ as observed from the camera with projection matrix $P \subset \mathbb{R}^{3\times 4}$.

To obtain $\bar{I}$, we iterate over all its pixels $\mathbf{u} \in [1..H] \times [1..W]$ and march with Emission-Absorption (EA) over the canonical ray $\hat{\mathbf{r}}_\mathbf{u}$ defined as the camera-space ray $\mathbf{r}_\mathbf{u} = P^{-1}\mathbf{u}$ in the rest-pose coordinates. Specifically, $\mathbf{r}_\mathbf{u}$ is first intersected with the posed outer boundary mesh $F(\hat{\mathcal{V}}^{\uparrow}, \beta, \theta)$, and then the intersection's barycentric coordinates are applied to the corresponding canonical mesh $\hat{\mathcal{V}}^{\uparrow}$ yielding a 3D point $\hat{\mathbf{x}}_\mathbf{u}^{\uparrow} \in \hat{\mathcal{N}}^{\epsilon}$ in the canonical neighborhood $\hat{\mathcal{N}}^{\epsilon}$. Repeating the same for the for the inner mesh $F(\hat{\mathcal{V}}^{\uparrow}, \beta, \theta)$ yields a second 3D point $\hat{\mathbf{x}}_\mathbf{u}^{\downarrow} \in \hat{\mathcal{N}}^{\epsilon}$. The two points $\hat{\mathbf{x}}_\mathbf{u}^{\uparrow}$ and $\hat{\mathbf{x}}_\mathbf{u}^{\downarrow}$ then define the canonical ray $\hat{\mathbf{r}}_\mathbf{u}$, over which we march with EA, accumulating the outputs of the radiance functions ψ_c and ψ_σ in the process, to render the final color of pixel $\mathbf{u}$ (details in the supplementary).

Note that our EA rendering differs from the duplex radiance fields [45], which instead leverage an MLP to directly map positional encodings of the two intersection points $\hat{\mathbf{x}}_\mathbf{u}^{\downarrow}$, $\hat{\mathbf{x}}_\mathbf{u}^{\uparrow}$ to a surface color.

3.2 Pose Initialization

Fitting a non-rigid shape to a monocular video is a challenging task and, as such, a trivial random initialization of the shape parameters (the weights of

MLPs ψ, χ) inevitably leads to a failure. We thus employ two fitting stages, where the first initializes parameters to ensure convergence of the second stage.

Root pose estimation. We observed that a suitable initialization of the root pose g^0, while initializing the rest of the parameters (limb angles θ^J, intrinsic deformation β, implicit MLP ψ) randomly, is sufficient to avoid the most common local minima, such as flipping of the dog body along its symmetry axes.

The goal of the initial fitting stage is thus to recover the root rigid transformations g_t^0 for $t \in [1, T]$ so that the perspective projection of the unposed canonical template $\hat{\mathcal{V}}$ into each camera P_t matches the depicted dog in frame t.

PnP with CSE. To this end, we leverage the CSE coordinate map of the SMAL mesh (Sec. 3.1). First, for each image I_t, a pre-trained CSE convolutional network annotates pixels $\mathbf{u}_t \in [1..H] \times [1..W]$ with their CSE embedding $e(\mathbf{u}_t)$. Then, we establish correspondences between each pixel $\mathbf{u}_t$ and the vertices $\hat{X}_{\mathrm{NN}^e(\mathbf{u}_t)} \in \hat{\mathcal{V}}$ on the template mesh $\hat{\mathcal{V}}$ by recovering the nearest neighbors $\mathrm{NN}^e(\mathbf{u}_t) := \arg\min_{1 \le k \le |\mathcal{V}|} \|e(\mathbf{u}_t) - e_k\|$ in the CSE embedding space. Given the set $\{(\mathbf{u}_t, \hat{X}_{\mathrm{NN}^e(\mathbf{u}_t)})\}$ of pixel-to-vertex correspondences, PnP-RANSAC [20] estimates the best camera P_t^{PnP} aligning the projections of the vertex points with their corresponding pixels.

In order to increase robustness to occasional failures of PnP caused by inaccurate CSE predictions, we employ a collective pose refinement that finds a single global rigid transformation $g^{\mathrm{PnP}} \in \mathbb{SE}(3)$ aligning the sequence of PnP-estimated cameras $(P_t^{\mathrm{PnP}})_{t=1}^T$ with the scene SfM cameras $(P_t^{\mathrm{SfM}})_{t=1}^T$ (the SfM cameras are detailed in the next section). The latter then initializes the root-rigid transformation of each frame, i.e., $\forall t : g_t^0 = g^{\mathrm{PnP}}$.

3.3 Shape Fitting

We now detail the second fine-fitting stage, which optimizes all shape parameters (β, θ_t, ψ) given the initial poses g^0.

Factoring the rigid pose. Even with near-perfect initialization of the root rigid pose, it is challenging to converge to a good solution when, as done in previous works [39,40], the rigid component $g_t = g_t^0 \in \mathbb{SE}(3)$ of the rendering camera P_t is solely represented with the root transformation g_t^0. Such optimization is challenging because g_t^0 has to represent the animal pose and also compensate for complex camera motions such as the jitter caused by unstable handling.

Instead, we factor the extrinsics $g_t = g_t^{\mathrm{cam}} \cdot g_t^0$ of each rendering camera P_t as a composition of the camera motion $g_t^{\mathrm{cam}} \in \mathbb{SE}(3)$ w.r.t. the rigid scene background and the motion $g_t^0 \in \mathbb{SE}(3)$ of the dog w.r.t. the background. Here, $g_t^{\mathrm{cam}} := g_t^{\mathrm{SfM}}$ is fixed to the SfM camera g_t^{SfM} estimated by COLMAP [41] which we empirically found to be very accurate. Offloading the camera estimate to SfM, we are then left with the simpler task of regressing the temporally-smooth rigid animal motion g_t^0.

CSE-guided fine shape fitting. The second fine fitting stage outputs all parameters (β, θ_t, ψ) by optimizing the shape predictor θ^{MLP} and the implicit shape

ψ using the following loss function:

$$\sum_{t=1}^{T}(\mathcal{L}_t^{\text{cse}} + \mathcal{L}_t^{\text{kp}} + \mathcal{L}_t^{\text{photo}} + \mathcal{L}_t^{\text{mask}} + \mathcal{L}_t^{\text{reg}}). \tag{5}$$

The latter calculates frame-specific losses $\mathcal{L}_t^{\cdot}$ and sums them over all T training images. The next paragraphs detail each loss term.

(i) CSE-keypoint loss. Besides leveraging CSE for the pose initialization, we also guide the fitting process with the following CSE keypoint loss:

$$\mathcal{L}_t^{\text{cse}} = \sum_{\mathbf{u}_t \in M_t} \|\mathbf{u}_t - P_t F(\mathcal{V}, \beta, \theta_t)_{\text{NN}^e(\mathbf{u}_t)}\|, \tag{6}$$

integrating the reprojection distances between each foreground pixel $\mathbf{u}_t \in M_t$ and the corresponding 2D projections $P_t F(\mathcal{V}, \beta, \theta_t)_{\text{NN}^e(\mathbf{u}_t)}$ of $\mathbf{u}_t$'s CSE correspondence $\text{NN}^e(\mathbf{u}_t)$ on the mesh $F(\mathcal{V}, \theta_t, \beta)_{\text{NN}^e(\mathbf{u}_t)}$. The mesh is deformed with parameters $\theta_t = \theta^{\text{MLP}}(\tau(t))$ given by the pose predictor θ^{MLP} for time-stamp $\tau(t)$ of the frame t. Here, $M_t \in \{0,1\}^{H \times W}$ corresponds to the foreground segmentation extracted with a pre-trained segmenter [4].

(ii) Sparse-keypoint loss. We complement the main CSE-keypoint training signal with the standard sparse-keypoint loss $\mathcal{L}_t^{\text{kp}} = \sum_{k=1}^{N_J} \|h(I_t)_k - P_t J_k\|$, minimizing the distance between the projection $P_t J_k$ and the detection $h(I_t)_k$ of the j-th SMAL joint $J_k = F(\hat{\mathcal{J}}, \beta, \theta_t)_k$ in the articulation of the image I_t. Here $h(I_t) \in \mathbb{R}^{N_J \times 2}$ denotes the 2D keypoint detections from BARC's sparse dog keypoint detector [39]. We discovered that sparse keypoints improve fitting on thin body structures, such as paws, where the CSE detections are less accurate.

(iii) Photometric loss. To train the appearance predictor, and to provide an additional supervision for fine 3D fitting, we minimize the photometric loss:

$$\mathcal{L}_t^{\text{photo}} = \text{LPIPS}(M_t \odot I_t, \mathcal{R}(P_t, F(\mathcal{V}, \theta_t, \beta), \mathcal{F}, \psi^{\text{MLP}})), \tag{7}$$

comprising the Learned Perceptual Image Patch Similarity [60] (LPIPS) between the masked ground-truth image $M_t \odot I_t$ and the RGB render of the posed mesh $F(\mathcal{V}, \theta_t, \beta)$ colored by the implicit duplex-mesh MLP ψ^{MLP}.

(iv) Chamfer mask loss. Similar to [3, 39, 40], we minimize a silhouette loss $\mathcal{L}_t^{\text{mask}}$:

$$\mathcal{L}_t^{\text{mask}} = d_{\text{Chamfer}}(\{\mathbf{u}_t | \mathbf{u}_t \in M_t\}, \{P_t X_k | X_k \in F(\mathcal{V}, \theta_t, \beta)\}), \tag{8}$$

evaluating Chamfer distance between the set of occupied 2D pixels $\mathbf{u}_t \in M_t$ of the g. t. mask M_t, and the 2D projections of the posed-mesh vertices $F(\mathcal{V}, \theta_t, \beta)$.

(v) Shape regularizers. To avoid implausible mesh articulations, we also employ shape regularization loss $\mathcal{L}_t^{\text{reg}} = \mathcal{L}_t^{\text{arap}} + \mathcal{L}_t^{\text{edge}}$, comprising the As-Rigid-As-Possible (ARAP) regularizer $\mathcal{L}_t^{\text{arap}}$ [43], and the edge-length penalty $\mathcal{L}_t^{\text{edge}}$ [16] enforcing local rigidity of the posed template. See supplementary for details.

Fig. 5: Qualitative comparison. We note that, unlike template-based approaches, the reconstructed meshes from RAC are very far from the actual shape of a dog. To evaluate temporal consistency, please refer to the result videos.

Table 1: Average results on 50 sequences from COP3D and 39 sequences from APTv2. Quality of pose estimates is measured by IoU, IoUw5, err_{track}; appearance quality is measured by PSNR, PSNRw5, LPIPS. Note that BARC and BITE only estimate pose and hence we cannot evaluate appearance quality, indicated by '×'.

Dataset	CoP3D [42]					APTv2 [56]
	IoU ↑	IoUw5 ↑	PSNR ↑	PSNRw5 ↑	LPIPS ↓	err_{track} ↓
BARC [39]	0.75	0.47	×	×	×	0.047
BITE [40]	0.81	0.59	×	×	×	0.047
RAC [55]	0.76	0.52	21.86	17.51	0.164	0.093
Ours	**0.84**	**0.79**	**22.12**	**19.40**	**0.041**	**0.035**

4 Experiments

Datasets. We evaluate all models on COP3D [42], an open-source dataset containing fly-around videos of pets annotated with cameras and segmentation masks. We select a subset of 50 dog videos with a variety of poses, movements and textures. Each test video contains 200 frames, which we split to the training and test sets by considering contiguous blocks of 15 frames as train, interleaved by blocks of 5 frames as test (we evaluate the impact of this protocol in Tab. 3). We evaluate all models at 256×256 resolution.

We also propose a tracking evaluation using APTv2 [56], a comprehensive dataset that includes videos featuring 30 distinct animal species in motion. Each video frame is annotated with 17 keypoints, which we use as ground-truth for quantitative evaluation. All videos in the dataset have a consistent length of 15 frames. We restrict to a subset of 39 sequences, where the video contains a single-instance of a dog.

Metrics. On COP3D [42], we report three evaluation metrics assessing the quality of the predicted shape and texture: (i) **IoU** reports the average over the intersection-over-union between the silhouette render of the posed shape S_t and the ground truth mask M_t computed for each frame t; (ii) **PSNR** computes the average peak-signal-to-noise ratio between the renders $\bar{I}_t$ of the fitted shape at time t, and the ground truth image I_t; and (iii) **LPIPS** [60] is a perceptual metric that measures the average visual similarity between the rendered images $\bar{I}_t$ of the fitted shape at time t, and the ground truth images I_t. This metric is particularly useful as it takes into account human visual perception and the structural information of the images, providing a more accurate measure of visual similarity compared to PSNR. Additionally, we report *worst 5%* variant **PSNRw5**, **IoUw5**. On APTv2 [56], we evaluate $\mathbf{err}_{\text{track}}$ measuring the tracking error of the annotated keypoints via the following protocol: (i) On the first frame, we pair each ground-truth keypoint kp_i^0 with a vertex $v_i \in \mathbb{R}^3$ on the mesh via the predicted pose. (ii) On the remaining frames, we compute *L2* distance between ground-truth keypoints and the projection of paired vertices.

Fig. 6: Texture transfer. Models optimized on a scene can be re-animated via the template articulations. We demonstrate shape and motion projection from one optimized scene to other textures.

Baselines. First, we compare to template-based reconstructors BARC [39] and BITE [40] which, similar to us, predict parameters of the SMAL model, but take as input only a single image (frame). Here, BITE extends BARC with a test-time iterative refinement. Secondly we compare to RAC [55], which is a template-free reconstructor with deformable shape and texture priors learned by observing various animals in videos. For fair comparison, we execute their shape/texture training pipeline on the original training set extended with our CoP3D videos. The quality of our textures can only be directly compared to RAC's model, which also includes texture, but we cannot compare them to BITE or BARC because they only output 3D shapes.

Results. Table 1 contains the results of our experiments on COP3D and APTv2. On COP3D, our method outperforms BARC,BITE and RAC in the quality of the predicted texture (LPIPS). In terms of IoU, we outperform BARC and RAC, and achieve similar performance to BITE. However, since our method is temporally-consistent, we significantly outperform BITE on worst 5% IoU. On APTv2 [56], our method outperforms BARC, BITE and RAC on the tracking evaluation. We note that RAC achieves significantly worst than all the other methods. We argue that this is caused by the poor quality of the rendered shape. We provide qualitative comparison in Fig. 5 and videos for visual evaluation.

Ablation study. To validate our design choices, we ablate individual components of our model and record the incurred changes in performance. *(i) Loss ablation.* In Tab. 2, we remove each loss term of Equation 5 and report the resulting PSNR/LPIPS/IoU. The results indicate a performance drop when any loss is removed, which confirms the merit of each loss. *(ii) Motion factorization.* We also demonstrate the benefits of our factorization of measured rigid motion

Table 2: Ablation on CoP3D reporting performance with various loss terms removed and without camera motion factorization ($g_t^{\text{cam}} = g_t^0$).

	w/o $\mathcal{L}^{\text{chamfer}}$	$\mathcal{L}^{\text{cse}}$	$\mathcal{L}^{\text{keypoint}}$	$\mathcal{L}^{\text{color}}$	$\mathcal{L}^{\text{arap}}$	$\mathcal{L}^{\text{edge}}$	$g_t^{\text{cam}} = g_t^0$	Ours
IoU ↑	0.70	0.81	0.80	0.81	0.81	0.83	0.72	**0.84**
PSNR ↑	20.65	20.89	21.54	21.62	21.61	21.88	19.12	**22.40**
LPIPS ↓	0.060	0.051	0.048	0.047	0.047	0.045	0.067	**0.038**

to the motion of the camera and the motion of the shape (Sec. 3.3). Specifically, we design an experiment where the rigid motion g_t of each rendering camera P_t is represented with the root rigid component g_t^0 of the SMAL deformation coefficients θ_t, i.e., $\forall t \in [1..T] g_t^{\text{cam}} = g_t^0$, ignoring the SfM estimate g_t^{SfM}. Results in Tab. 2 indicate that this leads to a significant decrease in performance across all metrics justifying our rigid motion factorization.

Varying frame-split difficulty. For COP3D evaluations [42], we follow the original train/test split protocol (contiguous blocks of 15 frames as train, interleaved by blocks of 5 frames as test). Removing excessive visible frames would render the problem unsolvable (e.g. it is impossible to guess the exact motion of a dog's legs given two boundary frames that are too far apart). Regardless, in Tab. 3 we evaluate the impact of reducing the supervision with two additional splits. Our method outperforms RAC on all splits, thus showing stronger robustness to weaker supervision.

Table 3: Evaluation with a varying train/test frame split on COP3D. Note that our method consistently beats RAC across a range of train/test splits.

Split Train/Test		IoU ↑	IoUw5 ↑	PSNR ↑	PSNRw5 ↑	LPIPS ↓
15/5	RAC [55]	0.76	0.52	21.86	17.51	0.164
	Ours	**0.84**	**0.79**	**22.12**	**19.40**	**0.041**
15/10	RAC [55]	0.71	0.4	21.54	16.46	0.175
	Ours	**0.82**	**0.69**	**22.07**	**18.54**	**0.048**
15/15	RAC [55]	0.66	0.37	20.67	15.38	0.196
	Ours	**0.81**	**0.63**	**21.62**	**17.81**	**0.056**

5 Discussion and Conclusions

Limitations. Our method, based on the CSE-SMAL template, is currently limited to reconstructing quadrupeds. Qualitative evaluations using COP3D show that thin body structures, particularly ears, are inaccurately reconstructed due to limitations in CSE predictions. Additionally, the expressiveness of the SMAL template restricts our reconstruction capabilities. For texture reconstruction, we rely solely on the input signal and do not inpaint unobserved areas, resulting in partial texturing from short videos where the animal is not fully visible.

Conclusion. In this paper, we proposed a novel method for generating textured animatable 3D mesh models from a casually captured monocular video of a dog. Our method augments the animatable SMAL mesh template with Continuous Surface Embeddings to setup a surface coordinate system which, in combination with a pretrained image-to-CSE predictor, allows to estimate accurate image-to-mesh correspondences that eventually lead to significantly more accurate fits. The better shape fitting then enables us to optimize an implicit opacity-color texture supported by a scaffold defined by the mesh in its rest pose. Our experiments reveal performance superior to existing template-free and template-based approaches on the challenging CoP3D dataset.

References

1. An, L., Ren, J., Yu, T., Hai, T., Jia, Y., Liu, Y.: Three-dimensional surface motion capture of multiple freely moving pigs using MAMMAL. Nature Communications (2023)
2. Badger, M., Wang, Y., Modh, A., Perkes, A., Kolotouros, N., Pfrommer, B., Schmidt, M., Daniilidis, K.: 3D bird reconstruction: a dataset, model, and shape recovery from a single view. In: Eur. Conf. Comput. Vis. (2020)
3. Biggs, B., Roddick, T., Fitzgibbon, A., Cipolla, R.: Creatures great and smal: Recovering the shape and motion of animals from video. In: Asian Conf. Comput. Vis. (2018)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Eur. Conf. Comput. Vis. (2020)
5. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3D generative adversarial networks. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
6. Chen, J., Zhang, Y., Kang, D., Zhe, X., Bao, L., Jia, X., Lu, H.: Animatable neural radiance fields from monocular RGB videos. ArXiv (2021)
7. Dong, Z., Chen, X., Yang, J., Black, M.J., Hilliges, O., Geiger, A.: AG3D: Learning to Generate 3D Avatars from 2D Image Collections. In: Int. Conf. Comput. Vis. (2023)
8. Ehsani, K., Bagherinezhad, H., Redmon, J., Mottaghi, R., Farhadi, A.: Who let the dogs out? modeling dog behavior from visual data. IEEE Conf. Comput. Vis. Pattern Recog. (2018)
9. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: Int. Conf. Comput. Vis. (2021)
10. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: Int. Conf. Comput. Vis. (2023)
11. Gu, C., Sun, C., Vijayanarasimhan, S., Pantofaru, C., Ross, D.A., Toderici, G., Li, Y., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: Ava: A video dataset of spatio-temporally localized atomic visual actions. IEEE Conf. Comput. Vis. Pattern Recog. (2017)
12. Guo, C., Jiang, T., Chen, X., Song, J., Hilliges, O.: Vid2avatar: 3D avatar reconstruction from videos in the wild via self-supervised scene decomposition. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
13. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. (2014)
14. Jakab, T., Li, R., Wu, S., Rupprecht, C., Vedaldi, A.: Farm3D: Learning articulated 3D animals by distilling 2D diffusion. Int. Conf. 3D Vis. (2023)
15. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: NeuMan: Neural human radiance field from a single video. In: Eur. Conf. Comput. Vis. (2022)
16. Kar, A., Tulsiani, S., Carreira, J., Malik, J.: Category-specific object reconstruction from a single image. IEEE Conf. Comput. Vis. Pattern Recog. (2014)
17. Kearney, S., Li, W., Parsons, M., Kim, K.I., Cosker, D.: RGBD-Dog: Predicting canine pose from rgbd sensors. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
18. Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L.: Novel dataset for fine-grained image categorization. In: First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition (2011)

19. Kokkinos, F., Kokkinos, I.: Learning monocular 3D reconstruction of articulated categories from motion. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
20. Lepetit, V., Moreno-Noguer, F., Fua, P.: EPnP: An accurate o(n) solution to the pnp problem. Int. J. Comput. Vis. (2009)
21. Lewis, J.P., Cordner, M., Fong, N.: Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. Association for Computing Machinery (2023)
22. Li, C., Lee, G.H.: Coarse-to-fine animal pose and shape estimation. ArXiv (2021)
23. Li, Z., Litvak, D., Li, R., Zhang, Y., Jakab, T., Rupprecht, C., Wu, S., Vedaldi, A., Wu, J.: Learning the 3D fauna of the web. ArXiv (2024)
24. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Eur. Conf. Comput. Vis. (2014)
25. Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3D reasoning. Int. Conf. Comput. Vis. (2019)
26. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. ACM Trans. Graph. (2019)
27. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and, M.J.B.: SMPL: A skinned multi- person linear model. ACM Trans. Graph. (2015)
28. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. In: 3DV (2017)
29. Melzi, S., Ren, J., Rodolà, E., Sharma, A., Wonka, P., Ovsjanikov, M.: ZoomOut: spectral upsampling for efficient shape correspondence. ACM Trans. Graph. (2019)
30. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Eur. Conf. Comput. Vis. (2020)
31. Neverova, N., Novotný, D., Vedaldi, A.: Continuous Surface Embeddings. In: Adv. Neural Inform. Process. Syst. (2020)
32. Novotny, D., Rocco, I., Sinha, S., Carlier, A., Kerchenbaum, G., Shapovalov, R., Smetanin, N., Neverova, N., Graham, B., Vedaldi, A.: Keytr: keypoint transporter for 3d reconstruction of deformable objects in videos. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
33. Osman, A.A.A., Bolkart, T., Black, M.J.: STAR: A sparse trained articulated human body regressor. In: Eur. Conf. Comput. Vis. (2020)
34. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: IEEE Conf. Comput. Vis. Pattern Recog. (2017)
35. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: Int. Conf. Comput. Vis. (2021)
36. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: Neural Radiance Fields for Dynamic Scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
37. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3D Deep Learning with PyTorch3D. ArXiv (2020)
38. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. ACM Trans. Graph. (2017)

39. Rueegg, N., Zuffi, S., Schindler, K., Black, M.J.: BARC: Learning to regress 3D dog shape from images by exploiting breed information. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
40. Rüegg, N., Tripathi, S., Schindler, K., Black, M.J., Zuffi, S.: BITE: Beyond priors for improved three-D dog pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
41. Schönberger, J.L., Frahm, J.M.: Structure-from-Motion Revisited. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016)
42. Sinha, S., Shapovalov, R., Reizenstein, J., Rocco, I., Neverova, N., Vedaldi, A., Novotny, D.: Common pets in 3D: Dynamic new-view synthesis of real-life deformable categories. IEEE Conf. Comput. Vis. Pattern Recog. (2023)
43. Sorkine, O., Cohen-Or, D., Lipman, Y., Alexa, M., Rössl, C., Seidel, H.P.: Laplacian Surface Editing. In: Proc. Eurographics (2004)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Adv. Neural Inform. Process. Syst. (2017)
45. Wan, Z., Richardt, C., Božič, A., Li, C., Rengarajan, V., Nam, S., Xiang, X., Li, T., Zhu, B., Ranjan, R., et al.: Learning neural duplex radiance fields for real-time view synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
46. Wang, Y., Kolotouros, N., Daniilidis, K., Badger, M.: Birds of a feather: Capturing avian shape models from images. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
47. Wiles, O., Gkioxari, G., Szeliski, R., Johnson, J.: Synsin: End-to-end view synthesis from a single image. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
48. Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., Wang, Y., Wang, Y.: AI challenger : A large-scale dataset for going deeper in image understanding. In: Int. Conf. Multimedia and Expo (2019)
49. Wu, S., Li, R., Jakab, T., Rupprecht, C., Vedaldi, A.: MagicPony: Learning articulated 3d animals in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
50. Xu, J., Zhang, Y., Peng, J.X., Ma, W., Jesslen, A., Ji, P., Hu, Q., Zhang, J., Liu, Q., Wang, J., Ji, W., Wang, C., Yuan, X., Kaushik, P., Zhang, G., Liu, J., Xie, Y., Cui, Y., Yuille, A.L., Kortylewski, A.: Animal3D: A comprehensive dataset of 3D animal pose and shape. Int. Conf. Comput. Vis. pp. 9065–9075 (2023)
51. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose++: Vision transformer for generic body pose estimation. IEEE Trans. Pattern Anal. Mach. Intell. (2024)
52. 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: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
53. Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Liu, C., Ramanan, D.: ViSER: Video-specific surface embeddings for articulated 3D shape reconstruction. In: Adv. Neural Inform. Process. Syst. (2021)
54. Yang, G., Vo, M., Neverova, N., Ramanan, D., Vedaldi, A., Joo, H.: BANMo: Building animatable 3D neural models from many casual videos. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
55. Yang, G., Wang, C., Reddy, N.D., Ramanan, D.: Reconstructing animatable categories from videos. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
56. Yang, Y., Deng, Y., Xu, Y., Zhang, J.: Aptv2: Benchmarking animal pose estimation and tracking with a large-scale dataset and beyond. ArXiv (2023)
57. Youwang, K., Ji-Yeon, K., Joo, K., Oh, T.H.: Unified 3D mesh recovery of humans and animals by learning animal exercise. Brit. Mach. Vis. Conf. (2021)

58. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: Neural Radiance Fields from One or Few Images. Eur. Conf. Comput. Vis. (2020)
59. Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: PyMAF-X: Towards well-aligned full-body model regression from monocular images. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
60. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018)
61. Zuffi, S., Kanazawa, A., Berger-Wolf, T., Black, M.J.: Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". Int. Conf. Comput. Vis. (2019)
62. Zuffi, S., Kanazawa, A., Jacobs, D.W., Black, M.J.: 3D menagerie: Modeling the 3D shape and pose of animals. IEEE Conf. Comput. Vis. Pattern Recog. (2016)