

Lions and Tigers and Bears: Capturing Non-Rigid, 3D, Articulated Shape from Images

Silvia Zuffi¹ Angjoo Kanazawa² Michael J. Black³

¹IMATI-CNR, Milan, Italy, ²University of California, Berkeley

³Max Planck Institute for Intelligent Systems, Tübingen, Germany

silvia@mi.imati.cnr.it, kanazawa@berkeley.edu, black@tuebingen.mpg.de

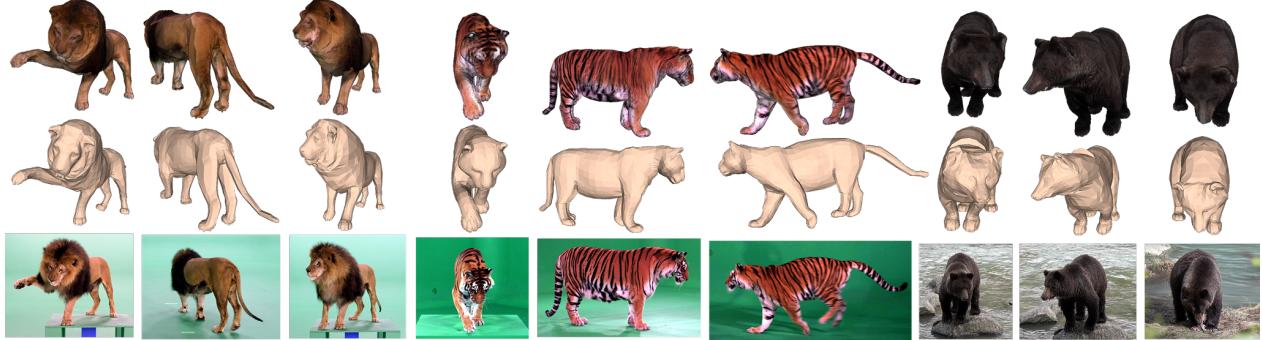


Figure 1: **Animals from Images.** We generate 3D textured articulated models of animals from images. Starting from an initial coarse shape obtained with the SMAL model [31], we refine the animal shape from multiple uncalibrated images and varying poses and capture a detailed texture map from multiple frames. We recover detailed 3D shapes for animal species that are in the SMAL shape space (like lions and tigers) and go beyond SMAL to capture new animals like bears.

Abstract

Animals are widespread in nature and the analysis of their shape and motion is important in many fields and industries. Modeling 3D animal shape, however, is difficult because the 3D scanning methods used to capture human shape are not applicable to wild animals or natural settings. Consequently, we propose a method to capture the detailed 3D shape of animals from images alone. The articulated and deformable nature of animals makes this problem extremely challenging, particularly in unconstrained environments with moving and uncalibrated cameras. To make this possible, we use a strong prior model of articulated animal shape that we fit to the image data. We then deform the animal shape in a canonical reference pose such that it matches image evidence when articulated and projected into multiple images. Our method extracts significantly more 3D shape detail than previous methods and is able to model new species, including the shape of an extinct animal, using only a few video frames. Additionally, the projected 3D shapes are accurate enough to facilitate the extraction of a realistic texture map from multiple frames.

1. Introduction

The study of animals is important not only for science, but for society as a whole. We need to learn more about animals in order to protect endangered species and to improve the quality of life for animals in captivity. We also need to share this knowledge with new generations who will increasingly acquire their awareness of nature from virtual worlds. Computer vision can play an important role by providing methods for the 3D capture, modeling, and tracking of animals. Unfortunately, few methods today support the capture of realistic animal shapes and textures like those in Figure 1.

In contrast, there are now mature tools for producing 3D articulated shape models of the human body that are realistic, differentiable, and computationally efficient. Such models support human pose, shape, and motion estimation from images and video as well as many applications of practical and commercial interest. In comparison to the human body, very little work has focused on modeling animals. The availability of 3D articulated shape models of animals, however, would open up many applications. Such models could be employed in markerless motion capture systems, removing intrusive markers and enabling the capture of wild

animals. Such tools would also support research in biomechanics, for example to understand the locomotion of different skeletal structures; in entertainment, where animal animation is still performed manually; in neuroscience, where tracking animals is fundamental for understanding behavior or for relating motion to brain activity; and in bio-inspired robotics, where understanding how animals move can help to design more efficient robots.

Despite extensive research on building human body shape models, it is not straightforward to extend these methods to model animal shape. The main reason is that human body models are built with the active collaboration of thousands of individuals who are captured with a 3D scanner in defined poses. This is clearly not practical for animals, in particular for wild species. Animal size and shape varies greatly, taking a scanner into the wild is impractical, getting them into a lab is a challenge, and they would need to be trained to adopt specific poses. Consequently, 3D training data is not available.

What is available in large quantities are images and videos of animals from nature photographs, animal documentaries, and webcams. Consequently, most previous work on modeling animal shape has focused on learning 3D models from images or video. Existing methods rely on manual intervention, as do we. Most previous approaches, however, are unable to extract realistic animal shapes. They either are limited to deformable but non-articulated shapes like dolphins or use part-based models where the parts are simple geometric primitives.

The key challenge in estimating animal shape from images is that they are articulated and deformable; in a video sequence, we observe the animal in different poses. With an unknown and moving camera, this problem is extremely complex. However, even though the animal may be in different poses, the underlying shape is the same. Thus we argue that the key is to explicitly disentangle the articulated structure of the animal from its shape. Doing so allows estimation of a consistent shape from images of the animal in many poses. In this work, given a collection of images, we simultaneously solve for the camera and the articulation parameters in each image, and a single shape of the animal in a canonical pose, such that when the shape is posed and projected into all the images, it matches the observed image data.

Given the complexity of the problem, we leverage a strong prior model of animal shape. In particular we use the articulated SMAL model, which captures a variety of animal shapes [31]. SMAL is learned from 3D scans of toy figurines spanning a range of animal species. We exploit two key aspects of the SMAL model: first, its vertex-based factorization of shape and articulation allows the estimation of a consistent shape given that this shape should be constant in all the images; second, its ability to approximate

animal shapes for novel species that are not present in the training set. While SMAL can generalize somewhat to new animal shapes, the reconstructed shape often lacks subject-specific detail. Thus, we use it as a starting point for refining the shape using image evidence. SMAL enables the reposing of the body into a canonical shape space, which allow us to combine information from multiple views and poses and improve the shape. Our method uses a few keypoints on the body and silhouettes of the animal in multiple images. The animal 3D shape, represented as a mesh, is deformed to match image evidence. In going beyond the SMAL model, the shape is poorly constrained by just a few views. Consequently, we regularize the deformation of the mesh from SMAL to constrain the final animal shape. We call the new model SMALR for “SMAL with Refinement” (pronounced “smaller”).

We show that the SMALR shape is visibly detailed, looks realistic, and resembles the specific animal instance (Fig. 1, middle). We show that the recovered shape is accurate enough to extract texture maps from multiple images and compose them into a full texture map for the animal. The textured results (Fig. 1, top) look quite realistic and may be used in animation, training-data generation, and tracking. A collection of 3D animal models are available at [1].

2. Previous

Our goal is to learn detailed 3D shapes of animals. To get sufficient real world data for this, we learn such models from unconstrained photographs and videos of animals in the wild. This presents significant technical challenges, which have not previously been addressed in the literature. Here we describe what has been done and where the previous approaches are insufficient for our task.

Avatars from 3D scans. There is extensive work on learning 3D, articulated, models of human bodies [2, 3, 17, 20]. This work begins with 3D surface scans of many people in many poses. A template mesh is then aligned to all scans, putting them into correspondence and enabling statistical modeling of shape and pose deformation. Unlike animals, humans are cooperative subjects, who will pose and stand still for scanning. Existing 3D scanners are not portable and can not easily be taken into the wild or even into zoos. Consequently, such methods do not immediately generalize to the problem of acquiring 3D animal shape.

Zuffi et al. [31] address this by using 3D scans of *toy* animals. They align a template mesh to a variety of animal species and build a statistical shape model similar to previous human models [20]. They fit this model to single-view images and show that the shape space generalizes somewhat to new species not seen in training.

Toys, however, are limited in number and realism. Not every animal is sufficiently popular for there to be realistic

toys depicting it. Consequently, here we go beyond previous work to use images and video, which are plentiful and capture the diversity of animal shapes.

In this work we also address the problem of extracting a texture map associated with the 3D shape. This has been done for humans in motion but only in controlled laboratory settings with a 4D scanner [5] or RGB-D sensor [4]. With the exception of Reinert et al. [23], there has been little work on extracting both animal shape and texture.

Rigid scenes from images. There is extensive work on recovering the shape of *rigid* scenes from sets of camera images and videos; we do not review this here. Classical multi-view stereo and structure from motion (SfM) methods assume all images are captured with the same camera and that the camera moves while the scene is static. More generally, *photo tourism* [25] and related methods take images from many different cameras and reconstruct a 3D scene. Again these methods assume the world is rigid.

Here we seek to do something similar but now the objects of interest are non-rigid and articulated; effectively *animal tourism*. Like photo tourism, we do not make assumptions that the camera is static, or that all images are captured by the same camera. This problem is significantly harder because of the non-rigid and articulated nature of animals. To make it possible, we constrain the shapes using a strong prior on animal shapes [31].

Performance capture. When multi-view stereo is applied to humans in movement, it is often called performance capture (e.g. [7, 10, 11, 27]). There are many methods that assume multiple static calibrated and synchronized cameras; again we do not summarize this extensive literature here. Typically such capture systems are installed indoors in environments with controlled backgrounds and lighting. A classical approach involves the extraction of silhouettes and the use of space carving [19] to extract a visual hull. Each silhouette provides a constraint on the body shape. We also use silhouettes but with a single moving camera and a moving animal. Consequently, in each frame, both the pose and the shape of the animal in the camera view is unknown and must be solved for.

There is less work on outdoor performance capture. Robertini et al. [24] take a scan of a person and deform it to fit multi-camera image evidence from calibrated outdoor cameras. The requirement of a pre-existing 3D scan of the actor to be captured is a significant limitation for animals.

Animals from images. Cashman and Fitzgibbon [8] learn a deformable model of animals from several images. They show an example of modeling a bear but with significantly lower fidelity than we show here (Fig. 2, left versus Fig. 1, right). Vincente and Agapito [29] use two views of a deforming object to recover a rough shape. Kanazawa et al. [18] learn separate animal models for cats and horses. They capture deformation using a variable stiffness model.



Figure 2: **Previous work.** Examples from [8] (left), [21] (middle) and [23] (right) .

All of these models are limited in their ability to model animal shape because they do not explicitly model articulation.

Ntouskos et al. [21] formulate a part-based model of animals. The shape of each part may be estimated from multiple segmented images and then assembled into a complete model. The final results retain the shapes of the parts and lack the realism of 3D scans (Fig. 2, middle).

Zuffi et al. [31] learn a parametric model of animal shape from figurines. Unlike the work above, this explicitly models articulation separately from animal shape. They learn a shape space that spans multiple animals and are able to fit the model to landmarks and silhouettes. The final fits are more realistic than previous methods but are overly smooth and generic. They are not adapted to the individual.

Animals from video. Video provides a potentially rich source of information about animal shape. Bregler et al. [6] show the estimation of the shape of a giraffe's neck and its modes of deformation from a video sequence using non-rigid structure from motion. They did not deal with articulation. Torresani et al. [28] estimate 3D structure from 2D point tracks. They show this for deformable objects like dolphins but the approach does not explicitly model articulation and does not estimate a 3D mesh. Ramanan et al. [22] build 2D models of animals from video using simple rectangular parts. Xu et al. [30] estimate animate gait cycles from a single image of multiple animals in different poses. Their method is only 2D and does not build a shape model of the animal. Similarly Favreau et al. [13] extract animal gait from video but not shape.

Most relevant here is the work of Reinert et al. [23] who show the extraction of a rough animal shape from video in terms of generalized cylinders. The animal shape is quite approximate due to the restriction of the parts to generalized cylinders. They also recover a texture map from one video frame but do not combine textures from multiple frames/views (Fig. 2, right).

Learning 3D models from images. There is recent work using CNNs to produce 3D models (typically voxel representations) from images [9, 16, 12]. So far these have focused on rigid objects like cars and chairs, where training data is plentiful. With animals there are few good 3D articulated models from which to train CNNs, which is why we aim to recover them from images and video.

In summary, our work occupies a unique position, com-

bining ideas from several fields to solve the challenging problem of uncalibrated 3D, articulated, animal shape estimation from several images.

3. Method

The SMAL model [31] we begin with can represent animals from 5 different families of quadrupeds: Felidae, Canidae, Equidae, Bovidae and Hippopotamidae. The shape of an animal is represented by a set of shape variables that define the vertex deformation applied to the model template to obtain a subject-specific shape. Formally, let β be a row vector of shape variables, then vertices of a subject-specific shape are computed as: $\mathbf{v}_{\text{shape}}(\beta) = \mathbf{v}_{\text{template},0} + B_s\beta$, where $\mathbf{v}_{\text{template},0}$ represents the vertices of the SMAL model template and B_s is a matrix of deformation vectors. Given a set of pose variables \mathbf{r} and global translation \mathbf{t} , the model generates 3D mesh vertices $\mathbf{v}(\beta, \mathbf{r}, \mathbf{t})$ in the desired pose with linear blend skinning.

In this work, as illustrated in Fig. 3, given a set of images of an animal annotated with landmarks and silhouettes, we obtain its 3D shape as follows. First, we align the SMAL model to the images, obtaining an estimate of the animal shape and its pose in each image. The shape estimate will be good for animals that are included in the SMAL shape model, and poor for animals that are not in the SMAL set. Second, we optimize a regularized deformation of the initial mesh to better match the silhouette and landmark annotations. Given an improved shape, we also update the pose in each image. Finally, we extract the texture of the animal from the images, resulting in a full, textured 3D model of the subject. We use 28 landmarks: 4 feet points, 4 knees points, 4 ankle points, 2 shoulder points, tail start, tail tip, neck, chin, 2 eye points, 2 nostril points, 2 cheeks points, 2 mouth points and 2 ear tip points (See Fig. 3, left).

SMAL Alignment to Images. Let $\{I^{(i)}\}$ be the set of N images of an animal obtained, for example, from frames of a video where the animal is seen in different poses and viewpoints. Let $\{S^{(i)}\}$ be the set of N silhouette images obtained by background subtraction or manual segmentation. Let $\{\mathbf{v}(\beta^{(i)}, \mathbf{r}^{(i)}, \mathbf{t}^{(i)})\}$ be the set of mesh vertices of N SMAL models, one for each frame, i , where the parameters are the shape variables $\beta^{(i)} \in \mathbb{R}^{20}$, global translation $\mathbf{t}^{(i)} = (t_x, t_y, t_z)$ and 3D pose parameters $\mathbf{r}^{(i)}$. \mathbf{r} is a concatenation of the relative rotation of 33 joints in a kinematic tree in axis-angle parameterization. The first 3 values capture the global rotation \mathbf{r}_0 . Let $\{\mathbf{K}^{(i)}\}$ be the set of n_K 2D landmarks $\mathbf{K}^{(i)} = \{\mathbf{k}_j^{(i)}\}_{j=1}^{n_K}$, manually annotated on each image. Each landmark is associated with a set of vertices on the 3D model, that we denote as $\{\mathbf{v}_{K,j}\}_{j=1}^{n_K}$. The number of model vertices associated with the j -th landmark is indicated as $n_{H(j)}$.

We model the camera with perspective projection, where

$\{\mathbf{c}(\mathbf{f}^{(i)}, \mathbf{r}_c^{(i)}, \mathbf{t}_c^{(i)})\}$ is the set of cameras defined by focal length $\mathbf{f} = (f_x, f_y)$, 3D rotation \mathbf{r}_c , and translation \mathbf{t}_c . We fix the extrinsic parameters to be at identity, and instead solve for the global pose \mathbf{r}_0 and translation \mathbf{t} of the animal. We set the principal point of the camera to be the image center. We also define shape priors and pose priors as in [31]. We use two shape priors: one defined for generic animals, and the family-specific shape prior of the SMAL model. We scale all images such that the maximum image dimension is 480 pixels. We first estimate the translation and the global rotation of the animal in each image by using a subset of the 2D landmarks corresponding to the animal torso. We initialize the pose of the N SMAL models to the mean pose in the pose prior and translation to zero. The shape variables β are initialized to zero or to the mean of the SMAL family-specific shape prior. For brevity, we drop arguments from functions when they are not being optimized over.

We first estimate the translation along the z axis. This is obtained as:

$$\hat{t}_z^{(i)} = f_x^{(i)} \text{median}\left(\frac{\|\mathbf{v}_{K,h} - \mathbf{v}_{K,l}\|_2}{\|\mathbf{k}_h^{(i)} - \mathbf{k}_l^{(i)}\|_2}\right), \quad (1)$$

where $\mathbf{v}_{K,h}$ is the mean of the $n_{H(h)}$ 3D model vertices associated with the h -th landmark, and (h, l) are indices of any combinations of two visible landmarks. Here the camera focus is set at $f_x = f_y = 1000$.

Then, we obtain estimates for translation and global rotation by solving an optimization problem:

$$\begin{aligned} \hat{\mathbf{t}}^{(i)}, \hat{\mathbf{r}}_0^{(i)} &= \arg \min_{\mathbf{t}^{(i)}, \mathbf{r}_0^{(i)}} \alpha_z(t_z^{(i)} - \hat{t}_z^{(i)}) + \\ &\sum_{j=1}^{\tilde{n}_K^{(i)}} \|\mathbf{k}_j^{(i)} - \frac{1}{n_{H(j)}} \sum_{h=1}^{n_{H(j)}} \Pi(\mathbf{v}(\mathbf{r}^{(i)}, \mathbf{t}^{(i)})_{K,j,h}, \mathbf{c}^{(i)})\|_2, \end{aligned} \quad (2)$$

where Π is the projection operator, and $\tilde{n}_K^{(i)}$ is the number of annotated landmarks on the i -th image.

Once we have obtained estimates for global rotation and translation for each image, we solve an optimization problem to estimate the articulated pose and shape parameters. We minimize an energy over all model parameters and the focal length on all images. Let $\Theta^{(i)} = (\beta^{(i)}, \mathbf{r}^{(i)}, \mathbf{t}^{(i)}, \mathbf{f}^{(i)})$ be the unknowns for the i -th image. Our objective is:

$$\begin{aligned} \hat{\Theta}^{(i)} &= \arg \min_{\beta, \mathbf{r}, \mathbf{t}, \mathbf{f}} \sum_{i=1}^N (E_{kp}(\Theta^{(i)}) + \\ &E_\beta^{(i)}(\beta) + E_{cam}(\mathbf{f}^{(i)}) + E_{sil}(\Theta^{(i)}) + \\ &E_{lim}(\mathbf{r}^{(i)}) + E_{pose}(\mathbf{r}^{(i)}) + E_{shape}(\beta^{(i)})). \end{aligned} \quad (3)$$

The term E_{kp} is the keypoint reprojection loss, defined as

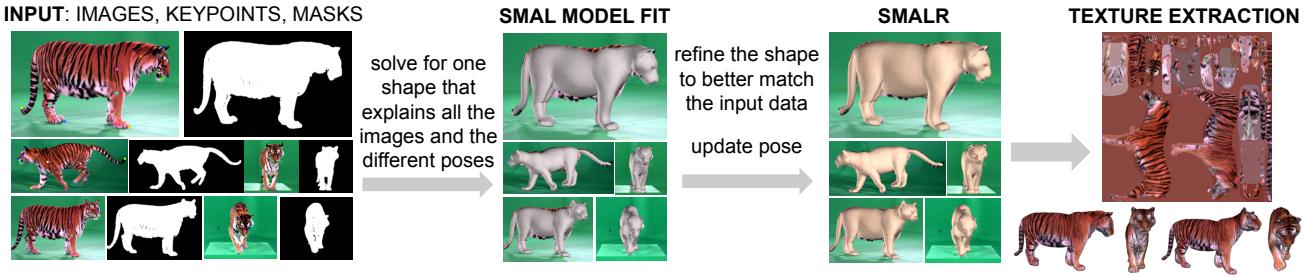


Figure 3: **Overview of the proposed pipeline.** We recover the instance-specific shape of an animal from multiple views, where the animal may be in a different pose in each image. We do this by first fitting SMAL to each image, which disentangles the pose (articulation) from the shape. Then we update the shape in a canonical pose, allowing it to deviate from SMAL, such that the recovered 3D shape, when articulated, better explains the image evidence in all views. We also update the pose to obtain a better fit to each image. The approach recovers a more accurate and detailed shape, which allows us to extract a coherent texture map from the images.

$$\text{in [31] as } E_{kp}(\Theta^{(i)}) = \alpha_{kp} \sum_{j=1}^{\tilde{n}_K^{(i)}} \rho(||\mathbf{k}_j^{(i)} - \frac{1}{n_{H(j)}} \sum_{h=1}^{n_{H(j)}} \Pi(\mathbf{v}(\beta^{(i)}, \mathbf{r}^{(i)}, \mathbf{t}^{(i)})_{K,j,h}, \mathbf{c}(\mathbf{f}^{(i)}))||_2), \quad (4)$$

Rather than estimate a single shape for all frames, the optimization is better behaved if we estimate different shapes for each frame but add a penalty on their difference; i.e. regularizing the shapes in all frames to be the same:

$$E_{\beta}^{(i)}(\beta) = \alpha_{\beta} |\beta^{(i-1)} - \beta^{(i)}| \quad \text{for } i > 1. \quad (5)$$

The term E_{cam} is a penalty for f_x and f_y to have the same value. We also add priors that encourage the camera focal length to be greater than 500 and the 3D shape to be in front of the camera. The terms E_{sil} , E_{lim} , E_{pose} , E_{shape} are defined as in [31]. Specifically, the silhouette term $E_{silh}(\Theta^{(i)})$ =

$$\alpha_{silh} \left(\sum_{\mathbf{x} \in \hat{S}^{(i)}} \mathcal{D}_S(\mathbf{x}) + \sum_{\mathbf{x} \in S^{(i)}} \rho \left(\min_{\hat{\mathbf{x}} \in \hat{S}^{(i)}} ||\mathbf{x} - \hat{\mathbf{x}}||_2 \right) \right), \quad (6)$$

where \hat{S} is the projected model silhouette, \mathcal{D}_S is the L2 distance transform field of the data silhouette such that if point \mathbf{x} is inside the silhouette $\mathcal{D}_S = 0$. The prior for joints limits, $E_{lim}(\mathbf{r}^{(i)})$ =

$$\alpha_{lim} (\max(\mathbf{r}^{(i)} - \mathbf{r}_{\max}, 0) + \max(\mathbf{r}_{\min} - \mathbf{r}^{(i)}, 0)). \quad (7)$$

\mathbf{r}_{\max} and \mathbf{r}_{\min} are the maximum and minimum range of values for each dimension of \mathbf{r} , respectively.

The pose prior E_{pose} is defined as the squared Mahalanobis distance using the mean and covariance of SMAL training poses. The shape prior E_{shape} is the squared Mahalanobis distance with zero mean and covariance given by

the SMAL shape space. When the animal family is known, we use the mean and the covariance of the training samples of the particular family.

At the end of the SMAL alignment we obtain estimates of pose, translation and shape for all the images. The penalty in Equation 5 encourages the shape variables to be the same for all images. We simply set the first shape to be the final shape $\hat{\beta}$.

SMALR Shape Recovery from Images. In this phase we capture more accurate 3D shape from images by estimating a deviation from the SMAL fit. For each animal, we define a vector of vertex displacements \mathbf{dv} that modifies the SMAL model as follows:

$$\mathbf{v}_{\text{shape}}(\mathbf{dv}) = \mathbf{v}_{\text{template},0} + B_s \hat{\beta} + \mathbf{dv}. \quad (8)$$

In this way we assign deformations to the 3D meshes that represent the animal in each image before the articulation is applied. With some abuse of notation, let E_{kp} and E_{sil} from above be redefined in the obvious way to take \mathbf{dv} as an argument, while keeping the pose, translation, camera and shape parameters fixed and set to the value of the previous optimization, $\hat{\Theta}^{(i)}$. Then, to estimate \mathbf{dv} , we minimize:

$$E_{opt}(\mathbf{dv}) = \sum_{i=1}^N (E_{kp}^{(i)}(\mathbf{dv}) + E_{sil}^{(i)}(\mathbf{dv})) + E_{arap}(\mathbf{dv}) + E_{sym}(\mathbf{dv}) + E_{lap}(\mathbf{dv}). \quad (9)$$

The terms E_{arap} , E_{sym} and E_{lap} are regularization terms, which are necessary to constrain the optimization. Namely, E_{arap} implements the as-rigid-as-possible deformation energy, which favors mesh deformations that are locally rigid rotations [26]. This regularization term is weaker for the head and mouth. The term E_{sym} favors the mesh



Figure 4: **SMALR**. Silhouette image (left), SMAL (middle) and SMALR (right). Here the silhouette image drives SMALR to produce a larger cheek on the animal’s left side compared with the SMAL initialization. The symmetry constraint also enlarges the right cheek, which is not constrained by the silhouette.

to be symmetric with respect to the main axis of the animal body. The term E_{lap} implements Laplacian smoothing [14]. Note that the Laplacian smoothing is defined over the displacement vector \mathbf{dv} and not on the vertices, therefore avoiding shrinking the mesh. The key idea here is that the data terms tries to move the vertices of the N models to fit the silhouette images and landmark positions. This happens under the constraint that the deformation should better explain *all* the silhouette images and landmarks positions jointly, while keeping the animal mesh smooth and symmetric. Figure 4 illustrates this with an example from a set of tiger images.

Once we have solved for \mathbf{dv} , we have the animal shape, $\mathbf{v}_{shape}(\mathbf{dv})$, and again perform the SMAL pose estimation step, keeping the shape fixed. The argument for this is that with a better shape, the pose estimate should improve.

Finally, after recovering detailed shapes of many animals, we take the new shapes and relearn the SMAL shape space, enabling it to capture a wider range of animal shapes. We do not illustrate this here.

Texture Recovery. In order to recover an animal’s texture from the images, we define a UV map of texture coordinates for the SMAL model. Given each image and corresponding estimated mesh, we define texture images and visibility weights for each texel. We combine the texture maps taking their weighted average. At this point we may have regions of the texture image that are not defined as the corresponding animal body surface is not visible in any of the images. We exploit the symmetry of the SMAL model to define a texture map corresponding to a symmetric mesh, where left and right side of the animal are swapped. Given the symmetric texture image, we can use it in two ways: we can assign to left/right corresponding texels the value of their average to recover a symmetric texture map, or we can fill in texels that are not defined with their corresponding ones in the symmetric texture image. We apply the first strategy to most of the animals, with the exception of animals with

stripes or large spots, which have large appearance differences between their left and right side. In the case a texel is undefined in both texture maps, we assign the median value of the colors of the body part to which it belongs. Figure 3 (right) illustrates the texture map obtained for the tiger.

4. Experiments

We evaluate SMALR quantitatively on a synthetic task, where we estimate the shape of an animal with known 3D shape, and qualitatively on challenging real images.

4.1. Synthetic Experiment

In this experiment we consider two scans of toy animals: a horse and a rhino. The horse species belongs to the Equidae family, which is represented in the SMAL model; the rhino species, on the contrary, is not represented in the SMAL model. Given a 3D scan, we can automatically generate sets of synthetic silhouette images and corresponding 2D keypoint annotations. However, if we just project the scans, this will not simulate the condition of different poses on different images that we assume in our work: we need to be able to animate them. To this end we define SMAL models where we replace the model template with the scans after registration to the SMAL template, and set the β shape variables to zero.

For each animal, we generate 4 sets of 5 silhouette images in random poses, where poses are sampled from a pose prior. In addition, we generate one set of five images with the animal in a canonical pose at fixed global poses: a side view, a top view, and three lateral views at different angles. We also generate the corresponding 2D landmarks annotations. We assume all the landmarks are visible.

We run our method on the synthetic images considering image sets of different size from 1 to 5. Figure 5 shows the silhouette images for the side views (left) with SMAL fit (middle) and SMALR fit (right).

We evaluate the mean distance between the estimated meshes and the ground-truth meshes for the estimates obtained with SMAL and with our SMALR method. Given that we do not fix the global translation and camera focus, the resulting meshes have a different scale with respect to the ground-truth meshes. In order to compute the mean distance between the results and ground-truth, we estimate a scale factor for the estimated meshes as the median ratio of vertex values for centered meshes.

Figure 5 shows the error in the estimation of the ground-truth meshes for the horse and rhino. The plot shows that with the method we can use a limited set of images with varying articulated pose and camera view to improve the estimate of the 3D animal shape.

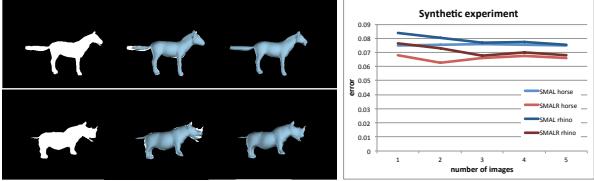


Figure 5: **Synthetic experiment.** (left) Example of synthetic silhouettes (1st col.), SMAL fit (2nd col.) and SMALR fit (3rd col.). These pictures refer to the case where two images are used, and the second view was from the top. (right) Error is average distance between estimated 3D mesh and ground truth mesh.

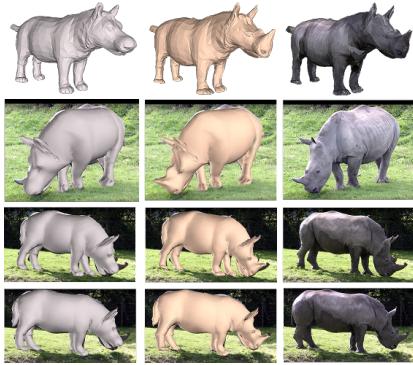


Figure 6: **Rhino.** We show the SMAL model fit (silver, left), the shape recovered by SMALR (gold, middle) and the recovered texture (top right) with input images (right). The SMAL model was not able to generate the horn, but this is recovered by SMALR.

4.2. Experiment on Real Images

We apply our method to generic animal images for the estimation of articulated 3D models with texture. We consider 14 animals: tiger, lion, bear, polar bear, panda, cheetah, wolf, dog, pig, horse, cow, cougar, and rhino. We also apply our method to an extinct animal, the Tasmanian tiger, for which a small amount of video exists. A summary picture of the animals considered is shown in Figure 8.

In order to be consistent with the SMAL model we only consider quadrupeds. Given that SMAL does not model articulated ears, we also consider animals or images with ears in a neutral position. In the case of the bears, we remove the tail from the SMAL model mesh.

In the case of the tiger, lion, cougar and cheetah, we consider frames extracted from different videos of the same animal captured against a green-screen background.¹ This enables the automatic extraction of accurate silhouettes. For all the other animals, except the pig, we extract frames from videos downloaded from YouTube. For the pig we use static

images of the same animal. The number of frames used for each animal varies from 3 to 7, depending on the availability of different poses and views. For the frames with arbitrary background we perform manual segmentation of the frames. All images are also annotated for landmark positions using an annotation tool. This process is quite simple and fast, as only a few images are necessary. For the cheetah we did not annotate all the face landmarks because the animal face is very small in the images. We optimize the SMAL model to approximate the animal shape and pose in each frame. It is important for the success of SMALR fitting that the pose is correctly estimated. We achieve good pose estimation in the majority of the cases. In a few frames the pose estimation failed because the poses were not well represented in our pose prior. In these cases we simply replaced the frames with different ones.

Detailed results for some of the animals considered are reported in Figures 1, 6, 7. For most of the animals considered, SMALR was able to recover a shape (gold color meshes) that better fits the input data compared with the SMAL initialization (silver colored meshes). In particular in the case of the rhino (Fig. 6), the horn is recovered, and the ears have a shape that better matches the picture. For the Tasmanian tiger (Fig. 7) our result produces a thinner snout compared with SMAL and slimmer hindquarters, which better resemble the animal.

Failure cases. It is important that the pose estimated by the SMAL fitting stage is correct, so that mismatches between silhouettes and landmark locations are imputable only to shape differences between SMAL and the actual animal shape. Further failure cases are when the silhouette images do not contain enough information to correctly recover shape details that characterize a specific species. Moreover, given in SMAL the ears are not articulated, and because we enforce mesh symmetry, errors occur if the animal has the ears in very different poses.

5. Conclusions

The 3D capture, modeling and tracking of animals has a wide range of applications where computer vision can play a major role by: making processes that are today performed mostly manually more automatic; allowing current studies to be performed at larger scales; stimulating new investigations in disciplines like biology, neuroscience, robotics; allowing fast creation of content for virtual and augmented worlds. All these tasks call for the availability of 3D articulated models of animal shape and appearance. In this work we have presented the first method that is able to capture high-quality 3D animal shape and appearance for different animals from images with little manual intervention. We do not assume multi-view capture nor calibrated cameras; our only assumption is that we can reliably estimate articulated pose. This is effectively a problem of non-

¹Data source: GreenScreenAnimals.com LLC.

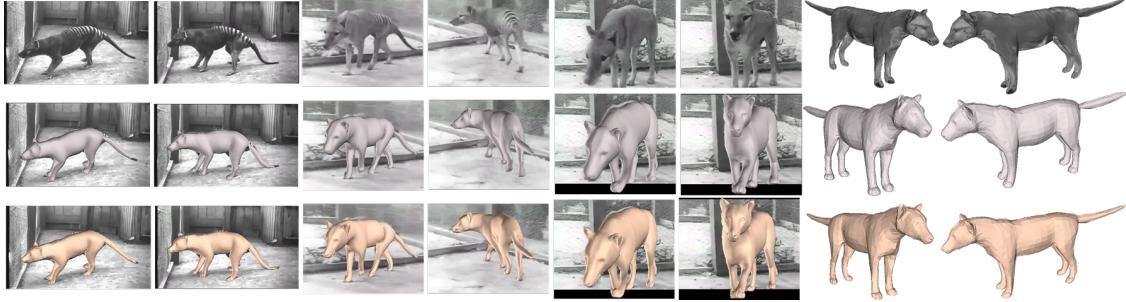


Figure 7: **Tasmanian tiger.** We show the SMAL model fit (silver, middle), the shape recovered by SMALR (bottom, gold) and the recovered texture (top right) with input images (top left). In this example the quality of the texture is compromised by the low quality of the video, captured in 1933 at the Hobart zoo, and from the absence of color information. The original animal was brown with black stripes. Here we are also not certain that the frames refer to the same individual.



Figure 8: **Results.** We show, for all the animals considered, the reconstructed 3D mesh (top), the mesh with texture applied (middle) and the image with the corresponding pose (bottom).

rigid structure from motion from unknown, moving, and uncalibrated cameras. As such it is extremely challenging. The key idea that makes this possible is that we exploit recent advances in modeling articulated animals to make the task of recovering accurate shape and texture from images tractable. We have shown the application of our method to a range of different animals. Here we focus on recovering intrinsic animal shape and assume generic pose-dependent deformations provided by SMAL. Future work should ex-

plore whether we can also learn pose-dependent deformations from images and video. Also, here we rely on manually segmented images and clicked feature points. Both of these processes could be automated with deep learning with sufficient training data. We plan to explore using our models to generate such training data synthetically. Finally we plan to exploit the SMALR models to capture the 3D motion animals in video sequences and expand the shape collection to a wider range of animal morphologies.

References

- [1] <http://smalr.is.tue.mpg.de>.
- [2] B. Allen, B. Curless, Z. Popović, and A. Hertzmann. Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis. In *Proc. of the ACM SIGGRAPH/Eurographics Symp. on Computer Animation*, SCA'06, pages 147–156, Aire-la-Ville, Switzerland, 2006.
- [3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape Completion and Animation of PEople. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 24(3):408–416, 2005.
- [4] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *International Conf. on Computer Vision (ICCV)*, pages 2300–2308, Dec. 2015.
- [5] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [6] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2:690–696, 2000.
- [7] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 326–339, Berlin, Heidelberg, 2010. Springer-Verlag.
- [8] T. J. Cashman and A. W. Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):232–244, Jan 2013.
- [9] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conf. on Computer Vision (ECCV)*, 2016.
- [10] S. Corazza, L. Mündermann, E. Gambaretto, G. Ferrigno, and T. P. Andriacchi. Markerless motion capture through visual hull, articulated icp and subject specific model generation. *International Journal of Computer Vision*, 87(1):156, Sep 2009.
- [11] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. *ACM Trans. Graph.*, 27(3):98:1–98:10, Aug. 2008.
- [12] A. Dosovitskiy, J. T. Springenberg, M. Tatarchenko, and T. Brox. Learning to generate chairs, tables and cars with convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):692–705, April 2017.
- [13] L. Favreau, L. Reveret, C. Depraz, and M.-P. Cani. Animal gaits from video. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 277–286. Eurographics Association, 2004.
- [14] D. A. Field. Laplacian smoothing and delaunay triangulations. *Communications in Applied Numerical Methods*, 4(6):709–712, 1988.
- [15] S. Geman and D. McClure. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, 52(4):5–21, 1987.
- [16] R. Girdhar, D. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In *European Conf. on Computer Vision (ECCV)*, 2016.
- [17] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H. Seidel. A statistical model of human pose and body shape. *Computer Graphics Forum*, 28(2):337–346, 2009.
- [18] A. Kanazawa, S. Kovalevsky, R. Basri, and D. W. Jacobs. Learning 3D deformation of animals from 2D images. In *Eurographics*, 2016.
- [19] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, Jul 2000.
- [20] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.
- [21] V. Ntouskos, M. Sanzari, B. Cafaro, F. Nardi, F. Natola, F. Pirri, and M. Ruiz. Component-wise modeling of articulated objects. In *International Conference on Computer Vision (ICCV)*, December 2015.
- [22] D. Ramanan, D. A. Forsyth, and K. Barnard. Building models of animals from video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(8):1319–1334, 2006.
- [23] B. Reinert, T. Ritschel, and H.-P. Seidel. Animated 3D creatures from single-view video by skeletal sketching. In *GI'16: Proc. of the 42nd Graphics Interface Conference*, 2016.
- [24] N. Robertini, D. Casas, H. Rhodin, H.-P. Seidel, and C. Theobalt. Model-based outdoor performance capture. In *Proceedings of the 2016 International Conference on 3D Vision (3DV 2016)*, 2016.
- [25] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, July 2006.
- [26] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing, Barcelona, Spain, July 4-6, 2007*, pages 109–116, 2007.
- [27] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Comput. Graph. Appl.*, 27(3):21–31, May 2007.
- [28] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):878–892, May 2008.
- [29] S. Vicente and L. Agapito. Balloon shapes: Reconstructing and deforming objects with volume from images. In *Conference on 3D Vision-3DV*, 2013.
- [30] X. Xu, L. Wan, X. Liu, T.-T. Wong, L. Wang, and C.-S. Leung. Animating animal motion from still. *ACM Transactions on Graphics (SIGGRAPH Asia 2008 issue)*, 27(5):117:1–117:8, December 2008.
- [31] S. Zuffi, A. Kanazawa, D. Jacobs, and M. J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.