VAREN: Very Accurate and Realistic Equine Network

Silvia Zuffi[1], Ylva Mellbin[2], Ci Li[3], Markus Hoeschle[4], Hedvig Kjellström[3,2], Senya Polikovsky[4],
Elin Hernlund[2], and Michael J. Black[4]

[1]IMATI-CNR, Milan, Italy, [2]SLU, Uppsala, Sweden, [3]KTH, Stockholm, Sweden,
[4]Max Planck Institute for Intelligent Systems, Tübingen, Germany

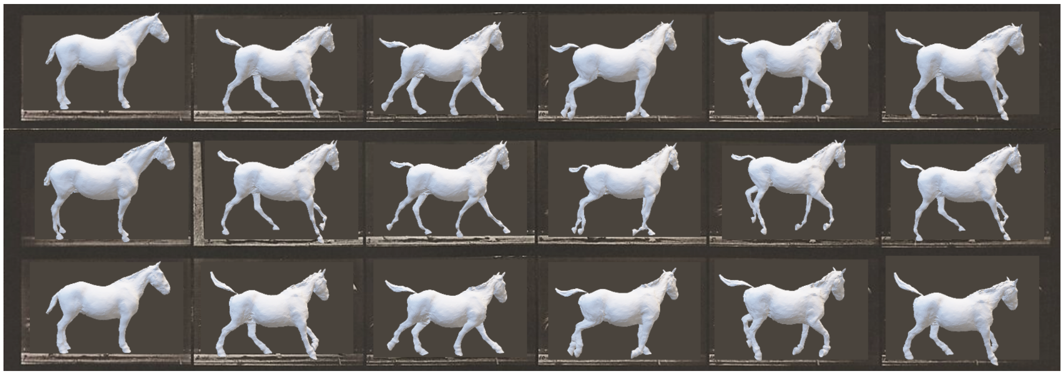

Figure 1. VAREN is learned from real horses and models pose-dependent deformations at the muscle level, enabling the generation of 3D horses with high realism. Inspired by Muybridge [24], we show a sequence of frames of horses in motion, generated by VAREN, where the first frame illustrates, for each row, the horse in the rest pose. We show three different shapes, obtained by sampling the VAREN model.

Abstract

Data-driven three-dimensional parametric shape models of the human body have gained enormous popularity both for the analysis of visual data and for the generation of synthetic humans. Following a similar approach for animals does not scale to the multitude of existing animal species, not to mention the difficulty of accessing subjects to scan in 3D. However, we argue that for domestic species of great importance, like the horse, it is a highly valuable investment to put effort into gathering a large dataset of real 3D scans, and learn a realistic 3D articulated shape model. We introduce VAREN, a novel 3D articulated parametric shape model learned from 3D scans of many real horses. VAREN bridges synthesis and analysis tasks, as the generated model instances have unprecedented realism, while being able to represent horses of different sizes and shapes. Differently from previous body models, VAREN has two resolutions, an anatomical skeleton, and interpretable, learned pose-dependent deformations, which are related to the body muscles. We show with experiments that this formulation has superior performance with respect to previous strategies for modeling pose-dependent deformations in the human body case, while also being more compact and allowing an analysis of the relationship between articulation and muscle deformation during articulated motion. The VAREN model and data are available at https://varen.is.tue.mpg.de.

1. Introduction

Horses are arguably the most valuable domestic animal and there is a large industry focused on their breeding, care, training, and use in sports. Buying a horse is a large investment, and keeping a horse requires resources and time. Although they are large animals, horses are delicate. The accumulated loads from training and competition frequently result in unrecoverable injuries in their skeletal or tendinous limb structures, ultimately leading to euthanasia. Consequently, among domestic animals, horses are widely studied from a behavioral and biomechanical perspective, with the aim of evaluating their performance, interpreting their state of pain, and preventing injuries. To facilitate such analysis using computer vision, our goal is to develop a highly accurate and detailed 3D model of horses that can be articulated, fit to data, and animated. Here we learn such a model

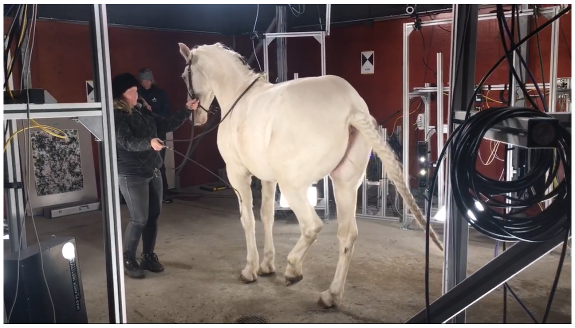

Figure 2. Dynamic 3D horse scanner. The horse is asked to lift one hind limb.

(see Fig. 1) using a novel formulation that captures pose-dependent muscle deformations. Such a model may facilitate 3D motion analysis in-the-wild, body shape estimation, diagnosis of illness and performance, horse-human interaction and behavior analysis, and generation of 3D synthetic horses for VFX, VR, AR, and gaming. In particular, accurate markerless motion capture of horses would provide a valuable new tool for the study of horse motion during challenging activities in natural environments.

For humans, there has been great progress on capturing their 3D shape and pose from images and video. This work often exploits the SMPL body model [22], a 3D mesh-based representation of the human body, controlled by 3D shape and pose parameters. SMPL provides strong priors over human shape and pose, enabling 3D pose reconstruction from ambiguous 2D data. Creating such a 3D model of horses is more challenging. While 3D articulated shape models of horses have been created in the past, they lack realism and expressiveness. In contrast, models like SMPL are learned from thousands of scans, and include an expressive shape space and a pose-dependent deformation model to represent how the body deforms under articulation. Unfortunately, obtaining high-quality 3D scans of horses in a variety of poses is challenging. To fill the gap, we need to (1) capture 3D scans of horses in motion and (2) formulate a new parametric 3D model of horses that represents important aspects of their physiology. In both cases, our solutions deviate from prior work on humans to accommodate the unique aspects of horses.

Dataset. We use a novel setup (3dMD Ltd.) to capture dynamic 3D scans of horses over time (see Fig. 2). Scanning a large number of horses presents technical and practical challenges, requiring a large cooperative effort between computer scientists and animal care experts. Our dataset includes 50 horses of different breeds and sizes, ranging from small ponies to large horses, with a height difference of more than 1.5 meters. This shape variation is significantly larger than with humans. To model pose-dependent shape changes, we need to capture the horses in a range of poses. Additionally, the animal is always managed by a care person who may occlude the horse during scanning, resulting in missing data. All these issues present novel challenges that we overcome. As a result, our dataset includes in total approximately 4000 cleaned up raw 3D scans, for which we provide accurately registered 3D meshes – training and testing split. Written consent was provided by all animal owners before the animals entered the study. The data collection procedure did not include any invasive technique, therefore no animal ethical approval was needed.

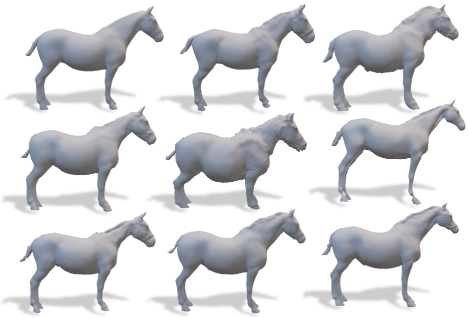

Figure 3. Samples from the VAREN model result in a diversity of horse shapes.

Model. Using this unique dataset, we train a new model named VAREN (pronounced Varenne[1] in French). VAREN is the first 3D articulated model of horses that is learned from real data and that models pose-dependent muscle deformations. Samples from the VAREN model are shown in a neutral pose in Fig. 3 and animated using motion capture data in Fig. 1. To model horses in motion, we must also capture the non-rigid deformations of the body that occur during movement. Previous work on humans models pose-dependent deformations that are learned from 3D scans and conditioned on the body articulation. In particular, STAR [25] models such deformations locally based on the distance from the nearby joints. Such methods do not explicitly model the deformation of the muscles. Scans of humans are typically acquired with some form of clothing (not fully naked) and most humans have a layer of subcutaneous adipose tissue that hides the musculature. Horses, however, are different. They generally carry less body fat than humans and their hair is short, making their musculature more directly observable. Modeling the visible muscle deformation is important because it relates directly to the health and performance of the animal. To that end, we propose a novel learned muscle-based deformation model. Specifically, we group the body surface points into regions that correspond to the superficial muscles, rather than based on their distance to the joints. Note that a single muscle can span more than one body part (see for example the back muscle in Fig. 4) and it is important to capture these long-range cor-

[1] Varenne is considered to be the best trotter of all time. No other trotter has won so many of the most important races in the world and set as many records as Varenne. Source: Wikipedia.

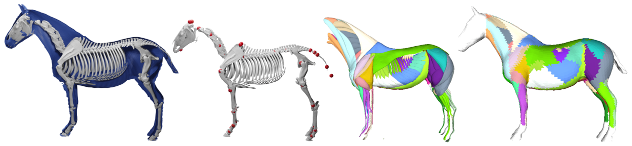

Figure 4. Skeleton and muscles. From left: anatomical skeleton manually aligned to the model template, selected joints, muscle structure and skin labels on the template. We consider 76 muscles. Head, tail and hooves have no muscle definitions.

relations. In VAREN, we define deformations per-muscle and learn the influence of body part articulation on the observable muscle deformation from the scan data. We find that, in comparison with previous formulations like those used in SMPL and STAR, our proposed muscle-inspired pose-dependent deformation model has superior accuracy, while also being more compact and interpretable. This is relevant for our high-resolution horse model. While we do not measure muscle activation directly, our formulation can be relevant for research on body pose using measurements of muscle activation [11]. A further novelty of our model, compared with previous work, is that we define the model skeleton, that is the location of the joints in the kinematic tree, based on real horse anatomy. Exploiting a 3D model of the horse skeleton, and with the help of an expert on horse biomechanics, we define the articulation points for the model. Previous human body models define the joint locations as a result of an optimization procedure, resulting in a per-model definition with poor correspondence to the true anatomical skeleton, in particular at the shoulders and hip joints [17]. These models are motivated by applications in computer graphics and vision, where, in many cases, a precise anatomical skeleton may not be needed. For a model like VAREN to be valuable to horse breeders, trainers, and veterinarians, it should provide information that is anatomically relevant. Moreover, an anatomical definition of the model joints, being model-independent, could facilitate comparison and motion data transfer between different models, and, eventually, species.

In summary, we make two key contributions. First, we provide a new dataset for the modeling of horse shape and pose. The quantity, resolution, variation, and quality of the VAREN data dwarfs any previous dataset of animal shape and pose, and opens up the possibility for the community to explore new questions in 3D shape representation. Second, we develop a new shape model that drives the observable deformation of the muscles based on the animal pose using an anatomical skeleton.

2. Related Work

We review prior work on learning human body models from data, as our problem and setting share similarities with methods proposed in this case. Then, we present the few 3D models for animals previously proposed. In addition, we provide an overview of methods for 3D animal reconstruction from visual data, in order to set the context in which we see our novel model applied.

Human Body Models. There is a long history of learning 3D mesh-based models of the human body from 3D scans [4–6, 10, 13, 22, 26, 35]. There is also recent work on implicit representations that we do not consider here, e.g. [3, 21, 23]. Most recent work is based on SMPL [22], a 3D body model that is based on vertex deformations defined by linear spaces. SMPL is fully data-driven: pose dependent deformations are defined as global representations and learned from data. One issue with SMPL is that the pose-dependent deformations are not local and capture spurious long-range correlations in the data. This is addressed by STAR [25], which adds locality to the joints to the pose-dependent deformations, weighting the outcome of global linear spaces with the distance from skeleton joints. While STAR conditions the pose deformation on the rough BMI of the body, none of the models above explicitly model the observable muscle deformation. VAREN addresses this for the first time. Also, unlike VAREN, the prior methods model articulation using a kinematic tree only loosely based on the human skeleton. A first direction into modeling the human skeleton is OSSO [17], who models the human body skeleton, followed by SKEL [18], which drives the pose of SMPL with a biomechanical skeletal model. SKEL, however, uses the default SMPL model deformations and, unlike VAREN does not train deformations that are driven by the skeletal pose.

Animal Body Models. The modeling of animal shape for computer vision applications has received less attention. Since 3D scanning of animals is challenging, research has mainly focused on learning from images; e.g. Ramanan et al. learn 2D articulated models [27], while Cashman and Fitzgibbon [9] adapt a rough mean shape using 2D image cues to reconstruct dolphins and bears in 3D. Zuffi et al. learn SMAL, a parametric 3D animal shape model, using 3D scans of toys [44]. Wang et al. learn the 3D shape variation of birds from images starting from a synthetic model [32]. Li et al. learn a horse-specific model similar to SMAL, using 3D scans of horse figurines [20]. Still, others learn to represent category-specific shapes as either meshes [39] or implicit surfaces [14], but do not decouple shape and pose. None of these methods learn 3D articulated models from real 3D scans, which contain noise. The multi-species SMAL model [44] contains an Equine class, but SMAL is learned from a limited number of toys, thus horses generated with the SMAL model do not exhibit a wide range of diverse shapes. The more recent hSMAL model is horse-specific [20]. While learned from a larger number of horses of different shapes, hSMAL is still learned from toys, and has a limited resolution. While useful for reconstructing horses from images and video, we find in our experiments that hSMAL is not expressive enough to accurately represent real horses with high detail. Being learned from rigid toys, both models do not address body deformations.

Animal 3D Reconstruction. The 3D reconstruction of animals from images and video follows a model-based or a model free approach. In the first case, an existing 3D shape model of the species of interest is available, and the method outputs the model parameters of 3D shape and pose. This approach is useful when dealing with challenging input modalities, like in-the-wild monocular images, or the goal is to estimate animal shape and pose for downstream analysis tasks. For example, Zuffi et al. address multi-species 3D reconstruction [45], Kanazawa et al. learn 3D deformations of animals [15], Badger et al. reconstruct 3D birds [7], while several methods estimate the shape and pose of dogs [8, 29, 30]. Our work fits in this category, but provides a 3D model with a higher level of realism than any previous methods. Model-free methods, in contrast, do not assume that a 3D model of the animal class is available, and the output of these methods is typically a 3D surface, in case of static input, or, eventually, a 3D animatable object in case of a video input. Works in this class include CMR [16] and DOVE [33], which reconstruct birds from images and ViSER [37], LASR [36], BANMo [38] and PPR [40], which reconstruct articulated 3D shapes from video, the latter incorporating physical constraints. Kokkinos and Kokkinos learn 3D reconstruction from video without using a shape model by relying only on a template [19]. MagicPony [34], LASSIE [41], Hi-LASSIE [42], ARTIC3D [43] learn 3D reconstruction from image collections.

3. Method

We learn the VAREN model in two stages. First, we learn an articulated parametric shape model from a set of scans, that we call *prototypes*, which are of different horses in a neutral pose, that is, not undergoing pose-dependent deformations. This gives us a model with a shape space and generic articulation. We then align, in an unsupervised way, this *static* model to the dynamic 3D scans in which the prototype horses perform different movements. From this data, namely the scans and the alignment parameters of 3D shape, pose and translation, the VAREN network learns a pose-dependent deformation model that improves the matching between the learned model and the scans. The novel deformation model is a function of the shape and pose parameters, such that, at test time, the network generates a shaped and posed horse, with realistic pose-dependent deformations, given the parameters. The pose-dependent deformation model is defined on the basis of anatomical structures that we incorporate into VAREN from a purchased realistic graphics (CG) model [1], which includes the body skeleton and the muscles. In the following, we detail the procedure to learn the horse model from the prototypes, the creation of the training dataset, and the definition of the VAREN network.

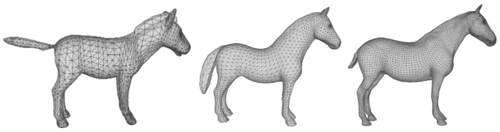

Figure 5. Comparison between the average equine model in SMAL (left), the hSMAL+ template (middle) and the VAREN template (right). Learned from real horses, the VAREN template has a more proportioned neck with respect to hSMAL+.

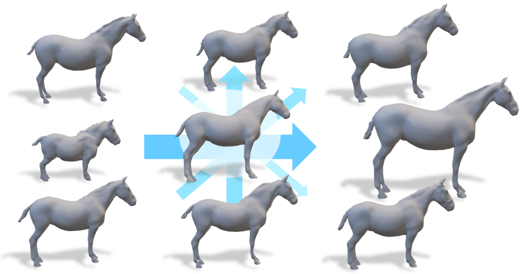

Figure 6. Visualization of the VAREN shape space. We capture small ponies, large breeds, and many in between, obtaining a shape space with variation in size, morphology and face features. The tail and mane are tied, and not considered as part of body shape change.

3.1. VAREN Horse Model

Alignment of the Prototype Scans. We follow a model-based approach to express the prototypes in a common topology, specifically, as the alignment model, we use hSMAL [20], a recently proposed articulated horse model (Fig. 5). The hSMAL model follows the formulation of SMPL, and is defined by a triangular mesh template $\mathbf{v}_t$, with n_V vertices, a matrix B of shape $3n_V \times n_B$ containing the n_B basis vectors of a linear shape deformation space, a joint regressor J_r that maps model vertices to a set of n_J joint locations, and a skinning weight matrix W. A horse is generated, given shape parameters β and pose parameters θ, by first deforming the template into an intrinsic shape $\mathbf{v}_s$, then applying Linear Blend Skinning (LBS) to rotate the body parts according to the given pose:

$$\begin{aligned} \mathbf{v}_s &= \mathbf{v}_t + B\beta^T \\ \mathbf{v} &= LBS(\mathbf{v}_s, \theta; W, J_r). \end{aligned} \tag{1}$$

While able to represent horses of different shapes, we found that the hSMAL model has, in our case, two limitations: a low resolution (the model has about $1.5K$ vertices) and a shape space that is not sufficiently expressive to accurately capture the shape of real horses. To overcome these limitations, we create an hSMAL model with an increased resolution (3647 vertices) and additional shape deformations.

We call this model hSMAL+. Furthermore, we observed that the purchased CG model [1], in addition to containing anatomical structures, has a more natural posture than hSMAL+. Therefore, we modify the rest pose of hSMAL+ to match the pose of the CG model. This is done by fitting hSMAL+ to the CG model and using the obtained mesh as the new template. In hSMAL+ we retain skinning weights and part segmentations from the original hSMAL model.

Additional Shape Deformations. We augment the hSMAL shape space by defining additional dimensions in the shape deformation matrix B (Eq. 1) to obtain a linear scaling of a set of body parts. Note that limb scaling has been applied before to animal models, but our formulation is novel, as previous work [8] applies scaling during the LBS process, while here we define scaling in the model shape space. Let j be a body part for which we want to add scaling as an additional dimension in the shape space. We define a new shape matrix B_1 of dimension $3n_V \times (n_B+1)$ by adding a new column to the shape matrix B:

$$\begin{aligned} B_1 &= B|B_j \\ B_j &= s_j \mathbf{v}_{j,c}, \end{aligned} \tag{2}$$

where $\mathbf{v}_{j,c}$ is a column vector corresponding to the template, but with non-zero values only for the coordinate c of the vertices of the part j we want to scale, and s_j is a scaling factor. Here j can also indicate a set of parts, for example, all the segments of the tail. Expanding B is not sufficient, as the deformation we obtain is expressed for the model template $\mathbf{v}_t$, while we need it to be applied to the part vertices of the intrinsic shape $\mathbf{v}_s$ (Eq. 1). Therefore, we add an iterative procedure that estimates a vertex shift d_v for the part vertices. Let J be the joints of the extended model with shape space B_1, J_0 be the joints of the original model, and J_s a vector of cumulative joint shifts initialized with zero values. In an iterative process that considers all the body joints from the root to the leaves of the kinematic tree, for each body part j, we compute the vertex shift as $d_v(j)$ = $J_s(j)$ + $J_0(j)$ - $J(j)$, add it to the part vertices, and then add $d_v(j)$ to the children of the part j in J_s, such that the joint shift is propagated through the kinematic tree. The set of body parts we consider for limb scaling are the ears, the legs and the tail, resulting in an additional set of 6 deformation vectors, as we group the left and right ears in a single part. The scaling factors in Equation 2 for these parts are 0.05 for the ears, where we scale the whole part, and 0.1 for the legs and tail, where we scale only the vertical axis. Note that the scaling factors in Equation 2 are applied to define the new shape space dimensions, but the scaling is then weighted with corresponding shape variables as for the original PCA dimensions.

Model-based Alignment. In order to align the hSMAL+ model to the prototypes, first, we manually annotate landmarks on the selected scans: for most of the prototypes we annotate 7 landmarks: the tip of the four hooves, the tip of the ears, and the tip of the tail. Note that this process, while manual, is applied to only one frame per horse, resulting in a very small fraction of the time invested to record a specific subject. The alignment procedure minimizes an energy with data terms that consist of the mesh-to-scan and scan-to-mesh distances between the scan and the model, and the distance between landmarks defined on the model and annotated on the scans. We add regularization terms for the β and the θ variables, set as the square of the elements of the β vector, and the $L1$ norm of the θ values, respectively. The latter formulation is as a consequence of the fact that in this stage we are aligning scans that are close to the model neutral pose by selection. The pose prior is stronger for the tail, which is often partially observed, and weaker for the ears, which might not be in a neutral pose. We optimize the energy with Chumpy [2], with an annealing schedule that decreases the weight of the pose prior and increases the weight of the data terms. After the model-based alignment, we increase the resolution of the hSMAL+ model (from 3647 to 13873 vertices) and perform a model-free alignment, where, to capture the fine-level details of the scans, we optimize over the vertices. In this case, we use two regularization terms: as-rigid-as-possible (ARAP) [31] regularization, and a coupling term between the 3D vertices and the model solution. Note that when increasing the model resolution, we do not alter the vertices that are inside the head (tongue and palate). Moreover, we keep the low-resolution vertices as the first 3647 vertices of the model, such that VAREN can be easily used at two resolutions.

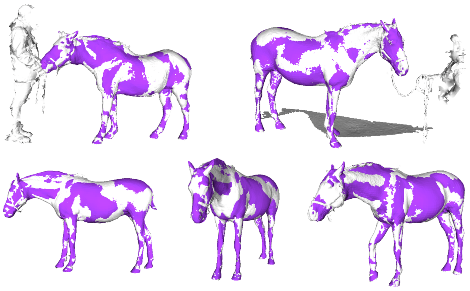

Figure 7. Model alignment to the scans. Examples of noisy (top) and clean (bottom) registrations. In white is the scan and in purple is the model registration.

VAREN Shape Space. The VAREN horse model shares the same formulation of SMPL and hSMAL expressed by Equation 1. The shape space of the model, indicated by the matrix B, is learned in the following way. Once the prototype scans have been aligned, they are all brought into the neutral pose of the alignment model (they are already close to this pose). The mean template shape is computed, giving an averaged prototype horse (Fig. 5), which is subtracted from the prototypes. Then Principal Component Analysis

(PCA) is applied, giving a linear shape space of skin deformations (Fig. 6). On the new template, we define the skeleton and muscle skin labels.

3.1.1 Anatomical Structures

In this section we describe how we incorporate anatomical structures, namely the skeleton and the muscles, into the VAREN model. Our goal is to define the model skeleton joints in correspondence with real joints, and characterize the model vertices according to the muscle they are closest to, as this association will be exploited in the pose-dependent deformation model. To this end, we exploit a realistic CG model of a horse [1] that includes a skeleton and muscles. We have already exploited the CG model to define the neutral pose for VAREN (Fig. 5), with the further advantage that now the VAREN template has the same pose of the skeleton of the CG model. However, given the body part proportions differ between the GC model and the VAREN template, we manually deform the individual skeleton bones to better match the template. We then select a set of joint locations on the skeleton that are anatomically relevant and closest to the hSMAL+ joints (Fig. 4). Finally, we re-compute the joint regressor of the VAREN model such that the anatomical joints are used. In addition to the existing hSMAL joints, we define two additional joints, corresponding to the left and right scapula. We found this necessary to fit the model to complex poses. A joint in correspondence to the top of the scapula models biomechanical principles for horses and dogs [12]. The muscles in the purchased CG model are grouped by large sections. We separate them into individual muscles, label them, and find the closest muscle for each vertex of the VAREN registration to the CG model. These labels are then transferred to the VAREN template. In this way, we obtain muscle labels for each vertex of the VAREN horse model (Fig. 4). Now that we have a model with anatomical joints and per-vertex muscle indices, first we re-register the prototypes to obtain the shape and pose parameters for the new model, then we learn pose-dependent deformations from the dynamic scans. The deformations are learned from an optimization network on a dataset of 3D scans aligned to the horse model. In the following, we describe the generation of the training set and the VAREN network.

3.2. Training Dataset

The training dataset is composed by a set of scans and corresponding alignment parameters. The scans are frames of a set of clips containing the same prototype horses used to learn the shape model performing different activities: standing still, moving the neck forward and toward a side, and moving the legs (see Fig. 2). As a pre-processing step, the scans are cleaned-up to remove the floor and the horse owner using a simple procedure based on point cloud clustering. We register the VAREN model to the dynamic sequences with an alignment procedure similar to the one used for the prototypes. The difference here is that we do not optimize for the horse shape, but keep the shape parameter fixed to the one computed with the VAREN prototype registration. For the first frame, we also initialize with the prototype pose. After the first, each frame is initialized with the previous solution, and the pose prior takes the form of the $L1$ distance from the previous pose. The alignment procedure does not require manual annotations. However, due to the presence of noisy frames, where the pre-processing clean-up failed, sometimes the alignment process gives a wrong result, and we need to annotate a few landmarks to correct such cases. We annotate about the 1% of the frames with about 4 landmarks. Moreover, we need to remove from the training set those frames where complete parts are missing. While we can deal with large holes in the scans, the alignment may return an incorrect solution if the entire head or a whole limb is missing. On the other hand, we allow ears to be missing in the scans, as the pose-dependent deformation model does not include them (see Fig. 4). Examples of registrations are shown in Fig. 7. The model registration can be used to further clean up and filter the data. We do this with a procedure that creates a set of bounding boxes around the registration to filter out the scan regions that are not pertinent to the horse (Fig. 8). We then compute the maximum scan-to-mesh distance to filter out scans with residual outliers. We retain, in total, about $3.6K$ training scans. Note that not all the prototype scans have been used to create the training set, as we leave out a few for testing.

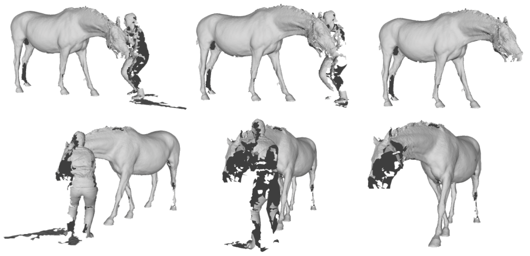

Figure 8. Post-alignment cleaning. We exploit the model registration to clean up the 3D scans. Left column: input scan, middle: cleaning with a single bounding box around the registration. Right: cleaning with a set of bounding boxes. The horse head is sometimes not captured due to the presence of a person.

3.3. VAREN Network

The pose-dependent deformations model is an additive term augmenting the intrinsic body shape (see Eq. 1):

$$\mathbf{v}_{s+d} = \mathbf{v}_t + B\beta^T + \mathbf{dv}_m(\theta, \beta), \quad (3)$$

where $\mathbf{dv}_m$ are the muscle deformations, obtained as:

$$\mathbf{dv}_{m,I_i}(\theta, \beta) = D_i(\beta_m), \quad (4)$$

where i is the muscle index, D_i is a decoder composed by a linear layer of dimension $(4(n_J{-}1){+}2){\times}3n_{I_i}$, with I_i the set of indices of the skin vertices associated with muscle i. The number of joints $n_J{=}38$, excluding the root, is multiplied by 4 as we use quaternions to represent pose. Equation 4 is applied for each muscle. The muscle deformation variable β_m is defined as:

$$\beta_m = A \circ W_m(\theta_{2,..n_J}, \beta_{1,2}), \tag{5}$$

where A is a selection matrix of dimension $n_M{\times}(4(n_J{-}1){+}2)$, with $n_M{=}76$ being the number of muscles. We initialize $A(i, 4k{:}4k{+}4) = 1$ if muscle i belongs to part j, with $k \in N(j)$, otherwise 0. Here $N(j)$ indicates the set composed by the part j and its neighbors. An advantage of our formulation is that each row of the product $A \circ W_m$ is now a weighting vector for the pose parameters indicating the influence between muscles and body parts, regardless of the pose, while the muscle deformation variable β_m indicates the strength of the muscle deformation. Both A and W_m are optimized during training. The last column of Fig. 9 shows the association between muscle deformation and body parts.

3.3.1 Network Training

During training, the network reads the training set, and stores the 3D pose, shape and translation estimated by the alignment procedure. Input scans are resampled at a fixed size of 20000 points to have a uniform size within a batch, with size 4. We store the alignment parameters in memory such that, during network optimization, pose and translation variables can be fine-tuned. However, we found in our experiments that fine-tuning the alignment variables does not significantly change the results, indicating that the training registrations are already of good quality. Here we present results for which we did not fine-tune the training data. We train using different losses: the Chamfer distance (from the PyTorch3D library [28]) between the generated horse mesh and the input scan, a regularization term, implemented as an edge length minimization loss, also from PyTorch3D, applied to the deformations $\mathbf{dv}_m$ and only to the mesh triangles on the muscle boundaries, and a regularization loss on the matrix A, to favor small values for body parts that are likely not to influence the muscles, namely the tail, feet, mouth and ears. This is implemented as the L_1 norm of the matrix entries for these parts. Weights for the losses are: $\alpha_{dist}{=}1e^3$, $\alpha_{reg}{=}10$, $\alpha_{bound}{=}100$, $\alpha_A{=}1e^3$.

4. Experiments

We apply the VAREN network to a set of held-out alignments. As a baseline, we consider the model without pose-dependent deformations. We also compare with defining pose-dependent deformations as in SMPL and STAR.

SMPL Pose-dependent Deformations. The SMPL model defines the pose-dependent deformations as a linear model over the pose vector of relative rotation matrices (see Eq. 9 in [22]). The pose feature vector is therefore of size $9(n_J{-}1)$, and the *pose blend shapes* is a matrix of shape $3n_V{\times}9(n_J{-}1)$. When training to learn the pose blend shapes, we use the Chamfer distance loss, the regularization loss applied to the deformations, and a loss on the L_1 norm of the pose blend shapes values. Weights for these losses are: $\alpha_{dist}{=}1e^3$, $\alpha_{reg}{=}100$, $\alpha_{reg,B_p}{=}1e^3$.

STAR Pose-dependent Deformations. The STAR model defines the pose-dependent deformations as a set of per-part linear models over the part and neighbors rotations, expressed with quaternions. The second value β_2 of the shape vector is added to the set of pose features, as it correlates with the human BMI index. Here, lacking this information for horses, we condition the deformation with the first two elements of the β vector, as in the VAREN shape space all the components capture morphological information (see Fig. 6). Each of the $n_J{-}1$ linear models has dimension $(4(n_{Jj}+1)+2){\times}3n_V$, where n_{Jj} is the number of neighbors of the part j. STAR defines the rotations relative to the human neutral pose, as the template pose in the STAR model is not neutral. This is not necessary for VAREN, where the template is in a natural rest pose. STAR also defines a set of activation weights, A, for each vertex, to weight the output of each linear model. Here, A is a matrix of shape $(n_J{-}1){\times}n_V$ that we initialize with the distance between vertices and joints. Following STAR, A is passed through a nonlinear rectifying function. During training, we use as losses the Chamfer distance and the regularization loss on the deformations. Weights for these losses are: $\alpha_{dist}{=}1e^3$, $\alpha_{reg}{=}100$. All networks are trained for 100 epochs, with a learning rate of $1e^{-5}$.

4.1. Test Datasets and Results

We consider two test sets. The **In-shape** test set, with 275 frames from 5 horses, includes the prototype horses used to train the shape deformation model, but not used for training the VAREN network. The **Out-shape** test set, 154 frames for 6 horses, includes subjects that have not been used at all. On the first set, we expect to obtain lower errors, and see only the effect of the pose-dependent deformation term, while the second set will provide a general performance assessment of the whole VAREN model. Results are reported in Table 1 and Table 2. We report errors in terms of average Chamfer distance, and of mesh-to-scan distance over a subset of model vertices obtained by removing the head, tail, ears, ankles and hooves. We do this to reduce the influence of the outliers and missing parts on the error scores. VAREN provides the best accuracy in comparison with the baseline and previous work. Figure 9 shows results for the Out-shape dataset. Notice how VAREN generates realistic

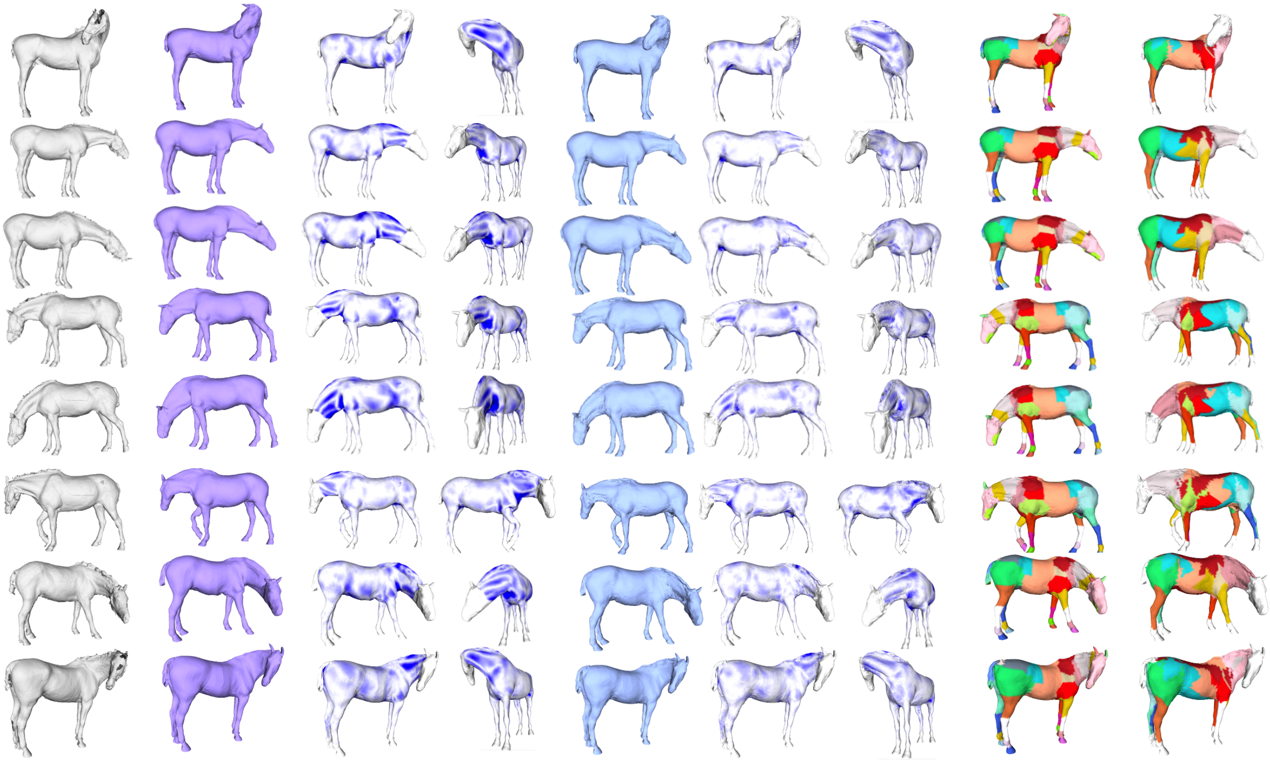

Figure 9. Results. From left: scan, baseline (purple), baseline mesh-to-scan distance (two views), VAREN (blue), VAREN mesh-to-scan distance (two views), VAREN with part colors, VAREN with muscle colored with the color of the part that gives the main contribution to the deformation, computed as the $argmax$ over body parts of the absolute value of the muscle deformation variable β_m (Eq. 5). For the mesh-to-scan visualization, distances are clipped at 4 cm, then normalized; darker regions indicate higher distance.

	Chamfer Distance		Mesh-to-Scan	
	mean	std	mean	std
Baseline	10.32	1.03	8.20	0.89
SMPL	9.72	0.98	7.06	0.72
STAR	9.51	0.67	6.99	0.63
VAREN (Our)	**9.37**	0.74	**6.35**	0.59

Table 1. Results on the In-shape testset. Baseline is the model without pose-dependent deformations. Errors in mm.

	Chamfer Distance		Mesh-to-Scan	
	mean	std	mean	std
hSMAL	21.92	3.02	17.01	2.51
Baseline	12.00	1.96	10.07	2.55
SMPL	11.38	1.66	8.84	2.01
STAR	11.03	1.57	8.36	1.77
VAREN (Our)	**10.88**	1.59	**7.78**	1.71

Table 2. Results on the Out-shape testset. hSMAL uses the hSMAL model for alignment. Baseline is the model without pose-dependent deformations. Errors in mm. Wilcoxon significance test for STAR and VAREN comparison: p-values are 0.04 and 3.3e-06 for Chamfer and mesh-to-scan distance, respectively.

deformations, in particular for the neck region. We also illustrate, for each muscle, the body part that most contributes to its deformation.

5. Conclusion

We introduced VAREN, the first parametric 3D model of horses learned from real data. In contrast to previous work, VAREN exploits a novel formulation that captures pose-dependent muscle deformations, resulting in better accuracy compared to state-of-the-art approaches, while also being more compact and connecting 3D pose to muscle deformations. By focusing on quality and anatomical realism, VAREN can support a wide set of AI applications in the equestrian world.

Acknowledgements. Silvia Zuffi is supported by PNRR project FAIR - Future AI Research (PE00000013), Spoke 8 - Pervasive AI (CUP H97G22000210007) under the NRRP MUR program by NextGenerationEU. We thank Minsoek Kim, Peter Kulits, Gökce Ergünfor, and Emil Breustedtfor for helping with pre-processing the scans. We also thank Dr. Zala Zgank, for assisting with data collection, and the horse owners for their time and for bringing their horses. **Disclosure**. MJB has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. MJB has financial interests in Amazon and Meshcapade GmbH. While MJB is a co-founder and Chief Scientist at Meshcapade, his research in this project was performed solely at, and funded solely by, the Max Planck Society.

References

[1] https://www.3dscanstore.com/3d-body-models/ecorche-3d-models/horse-ecorche-3d-model. 4, 5, 6

[2] http://chumpy.org. 5

[3] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5461–5470, 2021. 3

[4] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Trans. Graph.*, 22(3):587–594, 2003. 3

[5] Brett Allen, Brian Curless, Zoran Popović, and Aaron Hertzmann. Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 147–156, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association.

[6] 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. 3

[7] Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd G Pfrommer, Marc F Schmidt, and Kostas Daniilidis. 3d bird reconstruction: a dataset, model, and shape recovery from a single view. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–17, 2020. 4

[8] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3D animal reconstruction with expectation maximization in the loop. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020. 4, 5

[9] 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, 2013. 3

[10] Yinpeng Chen, Zicheng Liu, and Zhengyou Zhang. Tensor-based human body modeling. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 105–112, 2013. 3

[11] Mia Chiquier and Carl Vondrick. Muscles in action. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22091–22101, 2023. 3

[12] M.S. Fischer, K.E. Lilje, J. Lauströer, and A. Andikfar. *Dogs in Motion*. Kosmos, 2011. 6

[13] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.P. Seidel. A statistical model of human pose and body shape. *Computer Graphics Forum*, 28(2):337–346, 2009. 3

[14] Tomas Jakab, Ruining Li, Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Farm3D: Learning articulated 3D animals by distilling 2d diffusion. In *International Conference on 3D Vision (3DV)*, 2024. 3

[15] Angjoo Kanazawa, Shahar Kovalsky, Ronen Basri, and David Jacobs. Learning 3D deformation of animals from 2d images. *Comput. Graph. Forum*, 35(2):365–374, 2016. 4

[16] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 4

[17] Marilyn Keller, Silvia Zuffi, Michael J. Black, and Sergi Pujades. OSSO: Obtaining skeletal shape from outside. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 20492–20501, 2022. 3

[18] Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, Liu C. Karen, and Michael J. Black. From skin to skeleton: Towards biomechanically accurate 3d digital humans. *ACM ToG, Proc. SIGGRAPH Asia*, 2023. 3

[19] Filippos Kokkinos and Iasonas Kokkinos. Learning monocular 3D reconstruction of articulated categories from motion. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1737–1746, 2021. 4

[20] Ci Li, Nima Ghorbani, Sofia Broomé, Maheen Rashid, Michael J. Black, Elin Hernlund, Hedvig Kjellström, and Silvia Zuffi. hsmal: Detailed horse shape and pose reconstruction for motion pattern recognition. In *First CV4Animals Workshop, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 4

[21] Sandro Lombardi, Bangbang Yang, Tianxing Fan, Hujun Bao, Guofeng Zhang, Marc Pollefeys, and Zhaopeng Cui. Latenthuman: Shape-and-pose disentangled latent representation for human bodies. In *International Conference on 3D Vision (3DV)*, 2021. 3

[22] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2, 3, 7

[23] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[24] Eadweard Muybridge. *Animal Locomotion. An Electro-Photographic Investigation of Consecutive Phases of Animal Movements.* Philadelphia: University of Pennsylvania Press, 1887. 1

[25] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. STAR: Sparse trained articulated human body regressor. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 598–613, 2020. 2, 3

[26] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[27] Deva Ramanan, David A Forsyth, and Kobus Barnard. Building models of animals from video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(8):1319–1334, 2006. 3

[28] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia

Gkioxari. Accelerating 3D deep learning with PyTorch3D. *arXiv:2007.08501*, 2020. 7

[29] Nadine Rueegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. BARC: Learning to regress 3D dog shape from images by exploiting breed information. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3876–3884, 2022. 4

[30] Nadine Rüegg, Shashank Tripathi, Konrad Schindler, Michael J. Black, and Silvia Zuffi. BITE: Beyond priors for improved three-D dog pose estimation. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8867–8876, 2023. 4

[31] Olga Sorkine and Marc 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. 5

[32] Yufu Wang, Nikos Kolotouros, Kostas Daniilidis, and Marc Badger. Birds of a feather: Capturing avian shape models from images. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 14739–14749, 2021. 3

[33] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. DOVE: Learning deformable 3D objects by watching videos. *International Journal of Computer Vision (IJCV)*, 2023. 4

[34] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. MagicPony: Learning articulated 3D animals in the wild. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[35] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[36] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. LASR: Learning articulated shape reconstruction from a monocular video. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[37] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. ViSER: Video-specific surface embeddings for articulated 3D shape reconstruction. In *Neural Information Processing Systems (NeurIPS)*, 2021. 4

[38] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. BANMo: Building animatable 3D neural models from many casual videos. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2853–2863, 2022. 4

[39] Gengshan Yang, Chaoyang Wang, N Dinesh Reddy, and Deva Ramanan. Reconstructing animatable categories from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16995–17005, 2023. 3

[40] Gengshan Yang, Shuo Yang, John Z Zhang, Zachary Manchester, and Deva Ramanan. Ppr: Physically plausible reconstruction from monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3914–3924, 2023. 4

[41] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Lassie: Learning articulated shape from sparse image ensemble via 3D part discovery. In *Neural Information Processing Systems (NeurIPS)*, 2022. 4

[42] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[43] Chun-Han Yao, Amit Raj, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Artic3d: Learning robust articulated 3D shapes from noisy web image collections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 4

[44] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5524–5532, 2017. 3

[45] Silvia Zuffi, Angjoo Kanazawa, and Michael J. Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3955–3963, 2018. 4