VAREN: Very Accurate and Realistic Equine Network

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

\textsuperscript{1}IMATI-CNR, Milan, Italy, \textsuperscript{2}SLU, Uppsala, Sweden, \textsuperscript{3}KTH, Stockholm, Sweden, \textsuperscript{4}Max Planck Institute for Intelligent Systems, Tübingen, Germany

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 \url{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.
(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.

**Model.** Using this unique dataset, we train a new model named VAREN (pronounced Varenne¹ 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-

---

¹Varenne 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.
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 repre-
sent 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 anal-
ysis tasks. For example, Zuffi et al. address multi-species
3D reconstruction [45]. Kanazawa et al. learn 3D deforma-
tions 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 3D birds from images and
ViSER [37], LASR [36], BANMo [38] and PPR [40], which
reconstruct articulated 3D shapes from video, the latter in-
corporating 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 neu-
ral pose, that is, not undergoing pose-dependent deforma-
tions. 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 proto-
type 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 de-
formation model is a function of the shape and pose param-
eters, such that, at test time, the network generates a shaped
and posed horse, with realistic pose-dependent deforma-
tions, given the parameters. The pose-dependent deform-
ation 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

\begin{align}
\mathbf{v}_s &= \mathbf{v}_t + B\beta^T
\vspace{1em}
\mathbf{v} &= LBS(\mathbf{v}_s, \theta; W, J_T).
\end{align}

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.

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 net-
work.

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 \(v_t\), with
\(nv\) vertices, a matrix \(B\) of shape \(3nv \times nb\) containing the
\(nb\) basis vectors of a linear shape deformation space, a joint
regressor \(J_T\) that maps model vertices to a set of \(nj\) joint
locations, and a skinnning weight matrix \(W\). A horse is gen-
erated, given shape parameters \(\beta\) and pose parameters \(\theta\), 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:

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 lim-
itations, we create an hSMAL model with an increased res-
olution (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 $B$:

$$
B_1 = B|B_j
$$

$$
B_j = s_jv_{j,c},
$$

(2)

where $v_{j,c}$ is a column vector corresponding to the template, but with non-zero values only for all 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 $v_1$, while we need it to be applied to the part vertices of the intrinsic shape $v_s$ (Eq. 1). Therefore, we add an iterative procedure that estimates a vertex shift $d_{v_s}$ 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 $d_{v_s}(j) = J_s(j) + J_0(j) - J(j)$, add it to the part vertices, and then add $d_{v_s}(j)$ 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 $\beta$ and the $\theta$ variables, set as the square of the elements of the $\beta$ vector, and the $L1$ norm of the $\theta$ 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.

**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. The scans are cleaned-up to remove the floor and the horse head is sometimes not captured due to the presence of a person. 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.

### 3.3. VAREN Network

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

$$v_{s+d} = v_t + B\beta^T + dv_m(\theta, \beta),$$

where $dv_m$ are the muscle deformations, obtained as:

$$dv_{m,l}(\theta, \beta) = D_l(\beta_m),$$

where $m$ and $l$ index the muscle and frame, respectively.

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.

### 3.3. VAREN Network

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

$$v_{s+d} = v_t + B\beta^T + dv_m(\theta, \beta),$$

where $dv_m$ are the muscle deformations, obtained as:

$$dv_{m,l}(\theta, \beta) = D_l(\beta_m),$$

where $m$ and $l$ index the muscle and frame, respectively.

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.
where \( i \) is the muscle index, \( D_i \) is a decoder composed by a linear layer of dimension \((4(n_j-1)+2)\times 3n_J\), with \( I \) 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 \( \beta_m \) is defined as:

\[
\beta_m = A \circ W_m(\theta_{2:n_J}, \beta_{1:2}),
\]

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 \( \beta_m \) 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: \( \alpha_{dist}=1e^3 \), \( \alpha_{reg}=100 \), \( \alpha_{bound}=100 \), \( \alpha_A=1e^3 \).

### 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.

#### 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
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 Minsook 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


