

# BARC: Learning to Regress 3D Dog Shape from Images by Exploiting Breed Information

Nadine Rüegg<sup>1,2</sup>, Silvia Zuffi<sup>3</sup>, Konrad Schindler<sup>1</sup> and Michael J. Black<sup>2</sup>

<sup>1</sup>ETH Zürich, Switzerland

<sup>2</sup>Max Planck Institute for Intelligent Systems, Tübingen, Germany

<sup>3</sup>IMATI-CNR, Milan, Italy



Figure 1. *Monocular 3D shape and pose regression of 3D dogs from 2D images.* Since 3D training data is limited, BARC uses *breed* information at training time via triplet and classification losses to learn how to regress realistic 3D shapes at test time.

## Abstract

Our goal is to recover the 3D shape and pose of dogs from a single image. This is a challenging task because dogs exhibit a wide range of shapes and appearances, and are highly articulated. Recent work has proposed to directly regress the SMAL animal model, with additional limb scale parameters, from images. Our method, called BARC (Breed-Augmented Regression using Classification), goes beyond prior work in several important ways. First, we modify the SMAL shape space to be more appropriate for representing dog shape. But, even with a better shape model, the problem of regressing dog shape from an image is still challenging because we lack paired images with 3D ground truth. To compensate for the lack of paired data, we formulate novel losses that exploit information about dog breeds. In particular, we exploit the fact that dogs of the same breed have similar body shapes. We formulate a novel breed similarity loss consisting of two parts: One term encourages the shape of dogs from the same breed to be more similar than dogs of different breeds. The sec-

ond one, a breed classification loss, helps to produce recognizable breed-specific shapes. Through ablation studies, we find that our breed losses significantly improve shape accuracy over a baseline without them. We also compare BARC qualitatively to WLDO with a perceptual study and find that our approach produces dogs that are significantly more realistic. This work shows that a-priori information about genetic similarity can help to compensate for the lack of 3D training data. This concept may be applicable to other animal species or groups of species. Our code is publicly available for research purposes at <https://barc.is.tue.mpg.de/>.

## 1. Introduction

Learning to infer 3D models of articulated and non-rigid objects from 2D images is challenging. For the case of humans, recent methods leverage detailed parametric models of human body shape and pose, like SMPL [10]. Such models have been learned from thousands of high-resolution 3D scans of people in varied poses. This approach cannot be

replicated for most animal species because they are difficult, or even impossible, to scan in a controlled environment. Moreover, paired training data of animals with known 3D shape is even rarer. To make progress, we must leverage side information that can be easily obtained, yet constrains the task of 3D shape and pose estimation.

The 3D reconstruction of animal shape and pose has many real-life applications, ranging from biology and biomechanics to conservation. Specifically, the non-invasive capture of 3D body shape supports morphology and health-from-shape analysis. Markerless motion capture allows 3D motion analysis for animals that it is not possible to capture in a lab setting. Here we focus on dogs as a rich, representative, test case. Dogs exhibit a wide range of shapes, are non-rigid, and have complex articulation. Consequently, dogs are challenging and representative of many other animals.

Here, our goal is to learn to estimate a dog’s 3D shape and pose from a monocular, uncontrolled image. Given the lack of 3D training data, we train a regression network with 2D supervision, in the form of keypoints and silhouettes. With only such 2D information, the problem is, however, heavily under-constrained: many 3D shapes can explain the 2D image evidence equally well. To make the task well-posed, we need additional, prior information. Here we explore a novel source of a-priori knowledge: a dog’s shape is determined, in part, by its *breed*. Even a trained amateur can recognize the breed by looking at a dog’s shape (and appearance).

Dogs are a particularly interesting case to explore the role of breeds because of their large variety. Dogs have been domesticated and bred for a long time, for diverse purposes such as companionship, hunting, or herding, but also racing, pulling sleds, finding truffles, etc. Consequently, breeders have selected for a range of traits including body shape (as well as temperament, appearance, etc.) which has led to a large number of breeds with very different characteristics. A recent analysis of the dog genome illustrates the relationship between different breeds that exist today [5]. Breeds are grouped into *clades*, often with high shape similarity within a clade. Figure 2 shows a cladogram of 161 domestic dog breeds [5].

Here, we explore the use of genetic side information, in the form of breed labels, to train a regressor that infers 3D dog shape from 2D images. Specifically, we train a novel neural network called **BARC**, for “Breed-Augmented Regression using Classification.” We follow the approach of regressing a parametric 3D shape model directly from image pixels, which is common in human pose and shape estimation. Here, we use the SMAL animal model [32] to define the kinematic chain and mesh template. We extend SMAL in several ways to be a better foundation for learning about dog shape, this includes adding limb scale factors and

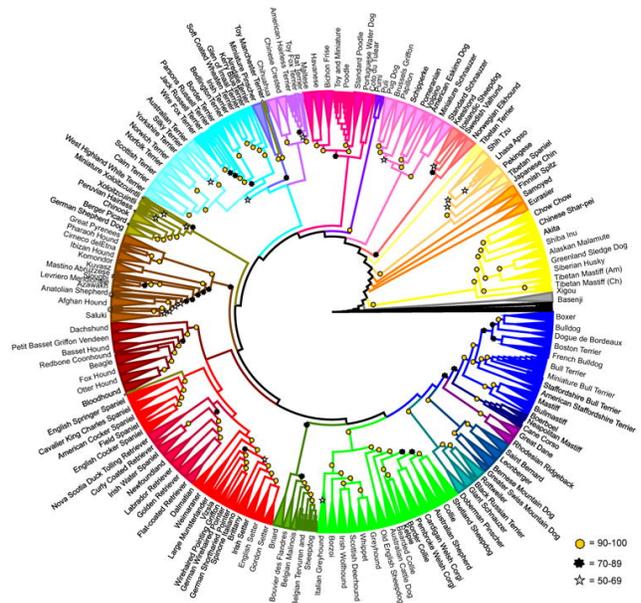


Figure 2. Cladogram of domestic dog breeds. The diagram represents clustering according to genetic similarity. Reproduced from [5].

extending its shape space with additional 3D dog shapes.

To solve the problem of estimating dog shape from images, we make several contributions. (1) We propose a novel neural network architecture to regress 3D dog shape and 3D pose from images. (2) To make training feasible from 2D silhouettes and keypoints, we exploit the fact that 2D images of the *same* breed should produce similar 3D shapes, while different breeds (mostly) have different shapes. With this assumption, we impose classification and triplet losses on the training images, which come with breed labels. (3) As a result, we learn a *breed-aware latent shape space*, in which we can identify breed clusters and relationships in agreement with the cladogram in Fig. 2. (4) Optionally, we show how to exploit 3D models, if available for some breeds.

Although we use one of the largest dog datasets in the literature, the large number of dog breeds (in our case 120) means there are only a few images per breed. One can interpret our method as learning a common shape manifold for all dogs (as not enough examples are available per breed), while using the breed labels to locally regularize it. To our knowledge, this is the first method that exploits breed information to regress the 3D shape of animals from images.

We train the network on the Stanford Extra (StanExt) [1, 9] training set, which we extend with eye, withers and throat keypoints, and test our approach on the StanExt test set. We evaluate on this dataset of 120 different dog breeds and show that we learn a latent shape space for dogs in which more closely related dogs are in closer proximity (Fig. 3).

Through ablation studies, we evaluate the impact of different types of breed information and find that each loss leads to a significant improvement in shape accuracy. We evaluate accuracy with standard 2D measures like PCK and IOU, but these do not necessarily reflect 3D accuracy. Consequently, we create a dataset of 3D dogs to evaluate shape of corresponding breeds. This allows a quantitative evaluation, and we significantly outperform the prior art (WLDO [1]). Finally, to evaluate shape estimates for in-the-wild images, we use a perceptual study to compare methods. We find that our final model is more realistic than ablated versions or WLDO.

## 2. Related Work

While many approaches focus on 3D reconstruction of humans from images, there is comparably little work on animal 3D pose and shape estimation. Animal reconstruction from images has been approached in two main ways: model-free and model-based.

**Model-free 3D Reconstruction.** These methods do not exploit an existing 3D shape model. Ntouskos et al. [14] create 3D animal shapes by assembling 3D primitives obtained by fitting manually segmented parts in multiple images of different animals from the same class. Vicente and Agapito [24] deform a template extracted from a reference image to fit a new image using keypoints and the silhouette, without addressing articulation. Kanazawa et al. [7] learn to regress 3D bird shape, given keypoints and silhouettes; birds exhibit rather limited articulation. Recent work obviates the need for 2D keypoints [6, 22, 26].

**Model-based 3D Reconstruction.** In one of the first 3D animal reconstruction methods from images, Cashman and Fitzgibbon [4] deform a 3D dolphin template, learning a low-dimensional deformation model from hand-clicked keypoints and manual segmentation. They also apply their method to a pigeon and a polar bear. A limitation of this approach is that articulation is not explicitly modeled. In contrast, Zuffi et al. introduce SMAL [32], a deformable 3D articulated quadruped animal model. Similar to the widely adopted human body model, SMPL [10], SMAL represents 3D articulated shapes with a low-dimensional linear shape space. Due to the lack of real 3D animal scans, SMAL is learned from scanned toy figurines of different quadruped species. Since dogs are not well represented by SMAL, Biggs et al. [1] extend the SMAL model by adding scale parameters for limb lengths. In [25], an articulated 3D model of birds is defined in terms of limb scale variations and used to learn shape from images; it is unclear whether this method easily extends to more complex animals.

The early work using SMAL uses an optimization-based approach to fit the model to image evidence [32] and to refine the animal shapes [31]. In other methods, Biggs et al. [2] show how to extract accurate animal shape and pose

from videos, while Kearney et al. [8] estimate dog shape and pose from RGBD-images. More relevant to BARC are learning-based methods that regress animal pose and shape directly. Biggs et al. [1] estimate dog pose and shape from single images by regressing pose and shape parameters of their model to training images of the StanExt dataset. Their initial pose prior is improved using expectation maximization with respect to fits of their model to the images. Zuffi et al. [30] regress a zebra SMAL model from images by exploiting a texture map and learn a shape space for the Grevy’s zebra. They train on synthetic data. In contrast to these methods, Sanakoyeu et al. [18] neither predict 3D directly from the image nor rely on sparsely annotated keypoints. Rather they show how to transfer DensePose from humans to a non-human primate. This approach does not recover 3D shape or pose.

**Supervision without 3D Ground Truth.** All 3D approaches rely on certain 2D features such as keypoints, segmentation masks or DensePose annotations as a supervision signal. Sometimes those 2D signals are used as an intermediate representation before the model is lifted to 3D. Mu et al. [12] exploit synthetic 3D data to predict 2D keypoints and a coarse body part segmentation map. They introduce a new dataset for animal 2D keypoint prediction and show how to transfer knowledge between domains, particularly from seen quadruped species to unseen ones. Still other work [6, 7, 22] encourages similarity between objects of similar shape, with small intra-class variability. They neither exploit breed information nor use contrastive learning to construct a structured latent space.

## 3. Approach

The present work explores how known breed information at training time can be leveraged to learn to regress a high-quality 3D model of dogs. To that end, we combine a parametric dog model with a neural network that maps images to model instances. In the following, we describe the model we use, the network architecture in which it is embedded, and the loss functions used to train the architecture, including the novel breed losses.

### 3.1. Dog Model

For the parametric representation of a dog’s shape and pose, we employ a variant of SMAL. We start from 41 scanned animal toy figurines of several different species (already used as part of the original SMAL model) as well as 3D *Unity* canine models in the animal equivalent of the canonical T-pose; i.e., standing with straight legs and tail pointing backwards. We purchased the same pack of *Unity* models<sup>1</sup> as was exploited by [1] to initialize their mixture

<sup>1</sup><https://assetstore.unity.com/packages/3d/characters/dog-big-pack-105660>

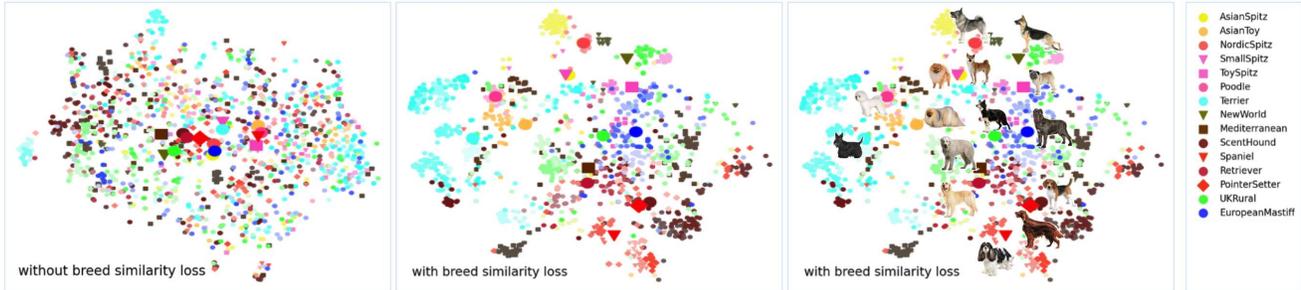


Figure 3. *Learned latent space.* t-SNE [23] visualization of the 64-dimensional latent shape variable for dogs in the test set. Large markers indicate average values within each of the clades in Fig. 2. *Left.* Latent space of the network trained without breed similarity loss. Note that the clade means are all near the population mean, indicating poor clustering. *Center and right:* With breed similarity loss. For each clade, colors with different saturation indicate different breeds within the clade.

of Gaussian shape prior and use them to relearn the SMAL shape space for our task. To that end, we fit a mesh with the same topology as SMAL (and WLDO) to the new dogs, add these to the original SMAL training set and recompute the mean shape and the PCA shape space. The resulting model differs from the original SMAL in three respects: (1) different input data; (2) reweighting of the inputs such that 50% of the total weight is assigned to dogs; and (3) rescaling of the meshes such that the torso always has length 1. We further adapt an idea from WLDO and extend the model with scaling parameters  $\kappa$  (where the actual scale is  $\exp(\kappa)$ ) for the limbs, plus an additional scale for the head length. The scaling is applied to the bone lengths, and propagated to the surface mesh via their corresponding linear blend skinning (LBS) weights. For compactness, we collect the PCA shape coefficients  $\beta_{pca}$  and limb scales  $\kappa$  into a shape vector  $\beta$ .

### 3.2. Architecture

Similar to [15, 28], we use separate shape and pose branches. Figure 4 shows the overall architecture of BARC, consisting of a joint stacked hourglass encoder, a shape branch, a pose branch, and a 3D prediction and reprojection module.

**Stacked Hourglass:** First, the input image is encoded and 2D keypoint heatmaps, as well as a segmentation map, are predicted with a pre-trained stacked hourglass network. 2D keypoint locations are extracted from the heatmaps with “numerical coordinate regression” (NCR, [13]). The segmentation map is encoded with a scheme similar to “basis point sets” (BPS, [16]) for 3D point cloud encoding. To our knowledge, we are the first to apply BSP in 2D. Compared to the full segmentation map, this encoding is lightweight, easy to compute for silhouettes, and has a similar format as the NCR keypoints. We find that, despite the reduction to a small number of sample points, the silhouette encoding still improves the 3D prediction over 2D keypoints alone.

**Shape-branch:** The input image and the predicted segmentation map are concatenated and fed to a ResNet34 that pre-

dicts a latent encoding  $z$  of the dog’s shape.  $z$  is decoded into both a breed (class) score and a vector of body shape coefficients  $\beta$ . We have experimented with different sub-networks between  $z$  and  $\beta$  and find that the breed similarity loss is most effective when the connection is as direct as possible, with only single, fully-connected layers between  $z$  and each of the shape vectors  $\kappa$  and  $\beta_{pca}$ . These shape coefficients are applied to the 3D dog template to obtain a shape, whose bone lengths are passed on to the pose branch.

**Pose-branch:** The predicted 2D keypoints, the BPS encoding of the silhouette and the bone lengths from the shape network form the input to estimate the dog’s 3D pose, its translation w.r.t. the camera coordinate system and the camera’s focal length. The pose is represented as a 6D rotation [29] for each joint, including a root rotation. Instead of predicting all rotations directly, we predict root rotation and a latent pose representation  $y$ . Following recent work on human pose estimation, we implement an invertible neural network (INN) that maps each latent variable  $y$  to a pose. This INN is used in the context of a normalizing flow pose prior trained on the RGBD-Dog dataset [8]. Similar to [27], we build this network consisting of Real-NVP blocks, but because of the smaller size of the RGBD-Dog dataset, as compared to AMASS [11], our network is much smaller than in previous work on human pose estimation. The aim of the INN is to map the distribution of 3D dog poses to a simple and tractable density function, i.e. a spherical multivariate Gaussian distribution. To train the pose prior, we exploit the RGBD-Dog dataset [8], which contains walking, trotting and jumping sequences, but no sitting or lying poses. Note that the INN is pretrained to serve as a pose prior and kept fixed during final network training.

**3D Prediction and Reprojection Module:** As a last step, BARC poses the model according to the predicted shape, pose and translation, and reprojects the keypoints and silhouette to image space, using the predicted focal length. To minimize the silhouette and keypoint reprojection errors, we employ the Pytorch3D differentiable renderer [17].

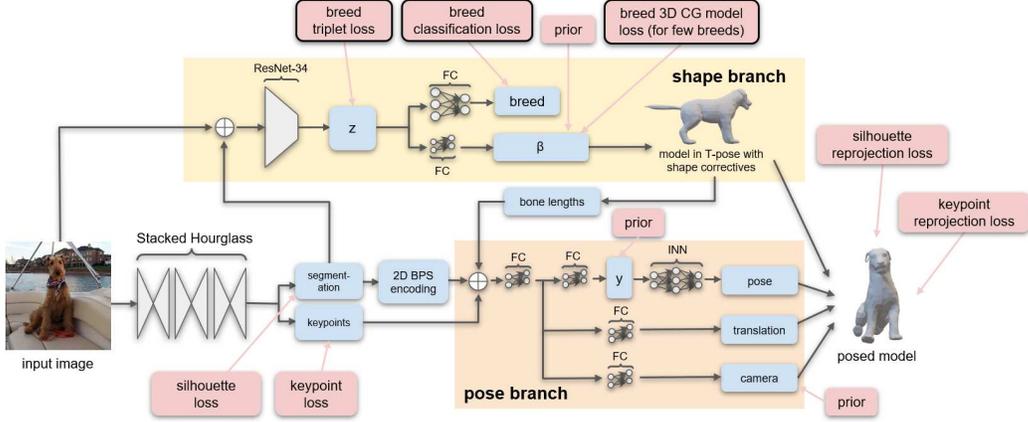


Figure 4. *BARC Architecture*. The model consists of a stacked hour glass network followed by two separate branches for shape and pose prediction. Pink boxes illustrate where losses are applied. The pink boxes with black boundaries are our new breed losses.

### 3.3. Training Procedure

The complexity of articulated, deformable 3D model fitting requires a number of different loss functions, as well as careful pretraining.

**Stacked Hourglass Pretraining:** The stacked hourglass is pretrained to predict keypoints and the segmentation map. The StanExt dog dataset [1] provides labels for both. The keypoint loss consists of two parts, a mean squared error (MSE) between the predicted and true heatmaps, and an L2-distance between the predicted and true keypoint coordinates. For the silhouette, we use the cross-entropy between ground truth and predicted masks. As usual for stacked hourglasses, we calculate the losses after every stage.

**Pose-branch Pretraining:** We use the same dataset (RGBD-Dog) that is used to train the pose prior to also pre-train the pose branch. We sample poses and random shapes and project them to a  $256 \times 256$  image with a random translation and focal length. The projected keypoints and silhouette serve as input to the network. MSE losses are used to penalize deviations between the predicted values and the ground truth. In addition, we use an MSE error between the predicted pose latent representation  $y$  and its ground truth.

**Main Training:** The stacked hourglass is kept fixed, while all other network parameters are jointly optimized. We point out that we do not have access to 3D ground truth, and, based on 2D keypoints, the true shape and pose is ambiguous. To regularize the solution, we therefore combine reprojection losses with suitable priors. These loss terms are described below.

### 3.4. Standard Losses

**Keypoint Reprojection Loss**  $L^{kp}$  is the weighted mean squared error between predicted  $k_n^{\text{pred}}$  and ground truth 2D

keypoint locations  $k_n^{\text{gt}}$ :

$$L^{kp} = \left( \sum_{n=1}^{N_{kp}} w_n d(k_n^{\text{pred}}, k_n^{\text{gt}})^2 \right) / \left( \sum_{n=1}^{N_{kp}} w_n \right), \quad (1)$$

where  $d(k_n^{\text{pred}}, k_n^{\text{gt}})$  is the 2D Euclidean distance between the predicted and ground truth location of the  $n$ -th keypoint. The weights,  $w_n$ , balance the influence of keypoints; see Sup. Mat.

**Silhouette Reprojection Loss**  $L^{\text{sil}}$  is the squared pixel error between the rendered  $s^{\text{pred}}$  and ground truth silhouette  $s^{\text{gt}}$ :

$$L^{\text{sil}} = \begin{cases} \sum_{x=1}^{256} \sum_{y=1}^{256} (s_{xy}^{\text{pred}} - s_{xy}^{\text{gt}})^2 & L^{\text{kp},m} < T \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

This is used only for images where the mean keypoint reprojection error  $L^{\text{kp},m}$  is below a threshold  $T$ .

**Shape Prior:** This is a weighted sum of two parts,  $L^{\text{sh}} = w_\beta L_\beta^{\text{sh}} + w_\kappa L_\kappa^{\text{sh}}$ . The first penalises deviations from a multivariate Gaussian with mean  $\mu_{\text{pca}}$  and covariance  $\Sigma_{\text{pca}}$ :

$$L_\beta^{\text{sh}} = (\beta_{\text{pca}} - \mu_{\text{pca}})^\top \Sigma_{\text{pca}}^{-1} (\beta_{\text{pca}} - \mu_{\text{pca}}). \quad (3)$$

Additionally, we penalise deviations from scale 1 with an element-wise squared loss on the scale factors  $\kappa$ ,

$$L_\kappa^{\text{sh}} = \sum_{i=1}^7 \kappa_i^2. \quad (4)$$

The shape prior loss is assigned a low weight and serves only to stabilise the shape against missing evidence.

**Pose Prior:**  $L^p$  penalises 3D poses that have low likelihood. Again, it consists of two terms, a normalizing flow pose prior as well as a regularization regarding lateral leg movements. The normalizing flow pose prior penalizes the negative log-likelihood of a given pose sample. Since the

learned latent representation  $y$  follows a multivariate normal distribution, the pose prior reduces to:

$$L_{\text{nf}}^{\text{p}} \propto y^{\top} y. \quad (5)$$

The normalizing flow prior is trained on the RGBD-Dog dataset which has a limited set of poses compared to the natural poses in the StanExt dataset. Consequently, with only this prior, the network can infer 3D poses where the legs move unnaturally sideways. Thus, we add a second term  $L_{\text{side}}^{\text{p}}$  that penalizes sideways poses of the three joints in each leg. The final pose prior is:

$$L^{\text{p}} = w_{\text{nf}} L_{\text{nf}}^{\text{p}} + w_{\text{side}} L_{\text{side}}^{\text{p}}, \quad (6)$$

with weights  $w_{\text{nf}}$  and  $w_{\text{side}}$ , the latter set to a low value.

**Camera Prior  $L^{\text{cam}}$ :** Since focal length  $f^{\text{pred}}$  is heavily correlated with depth (object-to-camera distance), we find it useful to penalise the squared deviation from a reasonably predefined target focal length  $f^{\text{target}}$ :

$$L^{\text{cam}} = (f^{\text{pred}} - f^{\text{target}})^2. \quad (7)$$

### 3.5. Novel Breed Losses

The losses described so far do not depend on the breed. To exploit breed labels for the training images, we introduce an additional breed triplet loss, as well as an auxiliary breed classification loss. We summarize those two losses as breed similarity loss. Given the dog meshes used during 3D model learning (Sec. 3.1) we moreover define a specific shape prior for those particular breeds.

**Breed Triplet Loss  $L_{\text{triplet}}^{\text{B}}$ :** Dogs of the same breed usually are somewhat similar in shape. However, this does not imply that there is no intra-class variation, nor that different breeds necessarily have dissimilar shape. Hence, we implement this with a triplet loss. We have experimented with different metric learning losses, but found that they all exhibit similar behaviour. Triplet losses are commonly used in person re-identification (ReID) methods, where the goal is to learn features that are discriminative for person identity [20, 21]. RingNet used a similar idea to learn 3D head shape from images without 3D supervision [19].

Applying the loss directly to the shape  $\beta$  does not work well. Shape changes along different principal directions may have different scales, moreover shape changes due to limb scaling are not orthogonal to the PCA coefficients  $\beta_{\text{pca}}$ . We find it better to apply the triplet loss to the latent encodings  $z$ . Given a batch with an anchor sample  $z_a$ , a positive sample  $z_p$  of the same breed and a negative sample  $z_n$  from a different breed, we calculate the triplet loss,  $L_{\text{triplet}}^{\text{B}} =$

$$\sum_{i=1}^{N_{\text{triplets}}} \max(d(z_{a,i}, z_{p,i}) - d(z_{n,i}, z_{a,i}) + m, 0), \quad (8)$$

where  $m$  is the margin and  $d$  denotes the distance between the two samples.

**Breed Classification Loss  $L_{\text{cs}}^{\text{B}}$ :** We further bias the estimation towards recognisable, breed-specific shapes with an auxiliary breed classification task, supervised with a standard cross-entropy loss on the breed labels:

$$L_{\text{cs}}^{\text{B}} = - \sum_{c=1}^{N_{\text{classes}}} y_{o,c} \log(p_{o,c}), \quad (9)$$

where  $p_{o,c}$  is the predicted probability that observation  $o$  is of class  $c$  and  $y$  is a binary indicator if label  $c$  is the correct class for observation  $o$ . The full similarity loss reads:

$$L_{\text{sim}}^{\text{B}} = w_{\text{triplet}} L_{\text{triplet}}^{\text{B}} + w_{\text{cs}} L_{\text{cs}}^{\text{B}}, \quad (10)$$

where  $w_{\text{triplet}}$  and  $w_{\text{cs}}$  are weights.

**3D Model Loss  $L_{\text{3D}}^{\text{B}}$ :** We have access to a small number of 3D dogs (*Unity* models) and a few 3D scans of toy figurines. These models encompass 11 of the 120 breeds in StanExt. For these breeds, we optionally enforce similarity between the prediction and the available 3D ground truth shape, via a component-wise loss on the shape coefficients  $\beta$ :

$$L_{\text{3D}}^{\text{B}} = (\beta_{\text{pca}}^{\text{pred}} - \beta_{\text{pca}}^{\text{breed}})^2 + (\kappa^{\text{pred}} - \kappa^{\text{breed}})^2. \quad (11)$$

## 4. Experiments

We evaluate our approach on the Stanford Extra Dog dataset (StanExt) [1]. StanExt provides labels for 20 keypoints, silhouette annotations and dog breed labels. We extend the 20 keypoints in the training set with withers, throat and eyes. These predictions are obtained by training a separate stacked hourglass on the Animal Pose dataset [3].

### 4.1. Evaluation Methods

**2D Reprojection Error:** In the absence of 3D ground truth, it is common to evaluate 3D shape and pose predictions in terms of reprojection errors in image space. We provide results for intersection over union (IoU) on the silhouette as well as percentage of correct keypoints (PCK).

**Perceptual Shape Evaluation:** Many implausible 3D shapes have low 2D reprojection errors, but for in-the-wild images we do not have access to ground-truth 3D shapes that would allow a meaningful comparison. Instead, we run a study to evaluate relative perceptual correctness where humans visually assess the 3D shapes regressed from in-the-wild images. Using Amazon Mechanical Turk (AMT), qualified workers judge which of two rendered 3D body shapes better corresponds to a query dog image. To focus the workers on shape, we render the dogs in the T-pose. For an example and details of the task see Sup. Mat.

**Breed Prototype Consistency:** Quantitative evaluation of 3D error for uncontrolled images is challenging. To address this, we exploit the fact that dogs of the same breed

Method	IoU	PCK @ 0.15				
		Avg	Legs	Tail	Ears	Face
3D-M	69.9	69.7	68.3	<b>68.0</b>	57.8	<b>93.7</b>
CGAS	63.5	28.6	30.7	34.5	25.9	24.1
WLDO	74.2	78.8	76.4	63.9	78.1	92.1
<b>Ours</b>	<b>75.7</b>	<b>83.7</b>	<b>83.9</b>	64.1	<b>82.8</b>	91.3

Table 1. *Comparison to SOTA*. Numbers for 3D-M [32], CGAS [2], WLDO [1] reproduced from [1].

have similar shapes. We define prototype shapes for several breeds with the help of toy figurines that are scanned, registered to the SMAL template, and reposed to the canonical T-pose. Then, for all StanExt images of the corresponding breeds, we regress their shape using various methods. These predictions are then also transferred to T-pose and aligned to the matching prototype with the Procrustes method. The vertex-to-vertex error and variance between the estimate and the prototype serve as indicators of how well a given prediction method captures the breed shape.

## 4.2. Comparison to Baselines

In terms of 2D error metrics (IoU and PCK) BARC outperforms prior art, i.e., WLDO [1], CGAS [2] and 3D-M [32]. Tab. 1 summarizes the results. In the perceptual comparison, BARC is also judged to represent the depicted dog better than its closest competitor WLDO, in an overwhelming 90.6% of all cases. See last line of Tab. 2. The marked gap in visual realism is evident in Fig. 5. More BARC results, for different breeds, are displayed in Fig. 6.

## 4.3. Ablation Study

Our key contribution is the addition of breed losses to improve 3D shape regression. To ablate the impact of individual loss terms, 2D errors are not meaningful, so we again report results in terms of relative perceptual correctness (Tab. 2) and in terms of consistency with the prototype breed shape (Tab. 3). We compare three versions of our method: (i) our network, trained without any breed losses; (ii) the same network with the breed similarity losses  $L_{sim}^B$ , i.e., classification and triplet loss; (iii) with all breed losses, including the 3D model loss  $L_{3D}^B$ . The results, on both metrics, show consistent improvements with the addition of each breed loss. In terms of perceptual agreement, the two parts of our loss have similar impact. The triplet and classification losses bring a clear improvement, even though they do not explicitly constrain 3D shape. Breed-specific 3D shape information can further improve the prediction, but may be difficult to collect at large scale. Note that adding 3D CG models as additional supervision leads to a small improvement (on average) across *all* breeds, even though they are only available for 11 out of 120 breeds. All differences in votes are highly significant ( $\chi^2$ -test,  $p < 0.0001$ ).



Figure 5. *Comparison to SOTA*. Qualitative comparison of BARC (left half) with WLDO [1] (right half). For each method we show input image, the 3D reconstruction projected on the input image, the 3D reconstruction, and a 90° rotated view.

Experiment Settings	AMT Results	
	Votes	Percentage
$L_{sim}^B$ vs. no breed losses	556 : 434	<b>56.2%</b> : 43.8%
$\{L_{sim}^B, L_{3D}^B\}$ vs. $L_{sim}^B$	678 : 522	<b>56.5%</b> : 43.5%
$\{L_{sim}^B, L_{3D}^B\}$ vs. WLDO	1033 : 107	<b>90.6%</b> : 9.4%

Table 2. *Perceptual Studies*. Ablation of breed losses and comparison with WLDO. See text.

We complement the perceptual study with a quantitative evaluation w.r.t. breed prototypes (Tab. 3). For 20 different breeds we evaluate WLDO, as well as our method without any breed losses, with only  $L_{sim}^B$ , and with both  $L_{sim}^B$  and  $L_{3D}^B$ . Already without breed information, our model outperforms WLDO by a clear margin in terms of 3D error, presumably due to technical choices like details of the dog model and network architecture, and the new pose prior. Adding the breed similarity losses decreases the error further. The additional 3D breed loss brings another reduction by a similar margin, which is consistent with the perceptual study. Again, all pairwise differences are highly significant (paired  $t$ -test,  $p < 0.0001$ ). Furthermore, the gains are consistent across breeds: For 19 out of 20 breeds we get the same order,  $WLDO > BARC_{nobreed} > BARC_{sim} > BARC_{sim+3D}$ .

**Breed Similarity Loss:** So far we have considered the two parts of the breed similarity loss  $L_{sim}^B$  in conjunction. To show the separate contributions of breed classification ( $L_{cs}^B$ ) and breed triplet affinity ( $L_{triplet}^B$ ) we evaluate consistency



Figure 6. *BARC* results. Each row shows the input image with the projected 3D shape. Below that is a rendering of the posed 3D shape.

Method	WLDO	BARC		
		no breed loss	$L_{sim}^B$	$\{L_{sim}^B, L_{3D}^B\}$
Error [m]	0.1155	0.0891	0.0805	<b>0.0716</b>

Table 3. *3D Shape Evaluation*. Average over 20 breeds.

with a breed prototype using different weights for the two terms, Tab. 4.  $w_{cs}^B$ ,  $w_{triplet}^B$ ,  $w_{3D}^B$  denote the weights for the classification, triplet and 3D-CG model loss, respectively. All other loss terms (regularizers, reprojection losses) remain fixed. Too a high weight on 3D shape similarity degrades the fit to the 2D image evidence (IoU, PCK). A good trade-off is  $w_{triplet}^B = 5$ .

Loss weights	$w_{cs}^B$	-	1	1	1	1	1
	$w_{triplet}^B$	-	-	5	10	5	10
	$w_{3D}^B$	-	-	-	-	1	1
Error [m]	0.089	0.082	0.081	0.074	0.072	<b>0.067</b>	

Table 4. *Ablation study*. Breed prototype consistency error for different setting

To make the influence of the breed information more tangible, we also visualize the effect of the breed similarity loss. Figure 3 shows a t-SNE visualization of the latent feature spaces learned by (left) a network without  $L_{sim}^B$  and (middle, right) an identical network trained with  $L_{sim}^B$ . The breed similarity pulls dogs of the same breed closer together in the latent space  $z$ , which is closely linked to the body shape parameters  $\beta$ . Different saturation levels of the same color indicate breeds within the clade. Even though the notion of clades is not imposed or made explicit anywhere in our network, breeds of the same clade tend to cluster. This suggests that not only within breeds, but also above breed level, shape knowledge can be transferred.

## 5. Conclusion

We present a method to reconstruct 3D pose and shape of dogs from images. Monocular 3D reconstruction of articulated objects is an unconstrained problem that requires strong priors on 3D shape and pose. We overcome the limitation of current 3D shape models of animals by training for model-based shape prediction with a novel breed-aware loss. We obtain state-of-the-art estimates of 3D dog shape and pose from images while also producing consistent, breed-specific 3D shape reconstructions. Our results outperform previous work metrically and perceptually. Combining visual appearance and genetic information through breed labels, we obtain a latent space that expresses relations between different breeds in accordance with recent studies on the dog breed genome. We believe this is the first work that combines breed information for learning to reconstruct 3D animal shape, and we hope it will be the basis of further investigation for other species.

**Limitations and Ethics.** BARC is limited by its shape space and is not able to go outside it. Given the high-quality regression results, future work should explore learning an improved shape space from images by exploiting breed constraints. We focused mainly on shape, but pose and motion are also important, and learning models of these from image data may be possible using our methods. Our research uses public image sources of dogs, and no animal experiments were conducted. While we focus on dogs, our method should be applicable to other animals and may eventually find positive uses in conservation, animal science, veterinary medicine, *etc.*

**Acknowledgements.** This research was supported by the Max Planck ETH Center for Learning Systems. A conflict of interest disclosure for Michael J. Black can be found here [https://files.is.tue.mpg.de/black/CoI\\_CVPR\\_2022.txt](https://files.is.tue.mpg.de/black/CoI_CVPR_2022.txt).

## References

- [1] Benjamin Biggs, Ollie Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out: 3D animal reconstruction with expectation maximization in the loop. In *ECCV*, pages 195–211, 2020. 2, 3, 5, 6, 7
- [2] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and SMAL: Recovering the shape and motion of animals from video. In *ACCV*, pages 3–19, 2018. 3, 7
- [3] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *ICCV*, pages 9498–9507, 2019. 6
- [4] Thomas J. Cashman and Andrew W. Fitzgibbon. What shape are dolphins? building 3D morphable models from 2D images. pages 232–244, 2013. 3
- [5] Parker H. G., Dreger D. L., Rimbault M., Davis B. W., Mullen A. B., Carpintero-Ramirez G., and Ostrander E. A. Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. *Cell Reports*, 4(19):697–708, 2017. 2
- [6] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoints without keypoints. In *ECCV*, pages 88–104, 2020. 3
- [7] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efron, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, pages 371–386, 2018. 3
- [8] Sinéad Kearney, Wenbin Li, Martin Parsons, Kwang In Kim, and Darren Cosker. RGBD-dog: Predicting canine pose from RGBD sensors. In *CVPR*, pages 8336–8345, 2020. 3, 4
- [9] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPR Workshops*, 2011. 2
- [10] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM ToG – SIGGRAPH Asia*, 34(6):248:1–248:16, 2015. 1, 3
- [11] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019. 4
- [12] Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille. Learning from synthetic animals. In *CVPR*, pages 12386–12395, 2020. 3
- [13] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018. 4
- [14] Valsamis Ntouskos, Marta Sanzari, Bruno Cafaro, Federico Nardi, Fabrizio Natola, Fiora Pirri, and Manuel Ruiz. Component-wise modeling of articulated objects. In *ICCV*, pages 2327–2335, 2015. 3
- [15] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, pages 459–468, 2018. 4
- [16] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *ICCV*, pages 4332–4341, 2019. 4
- [17] 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. 4
- [18] Artsiom Sanakoyeu, Vasil Khalidov, Maureen S McCarthy, Andrea Vedaldi, and Natalia Neverova. Transferring dense pose to proximal animal classes. In *CVPR*, pages 5233–5242, 2020. 3
- [19] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *CVPR*, pages 7763–7772, 2019. 6
- [20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 6
- [21] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014. 6
- [22] Shubham Tulsiani, Nilesh Kulkarni, and Abhinav Gupta. Implicit mesh reconstruction from unannotated image collections. *arXiv preprint arXiv:2007.08504*, 2020. 3
- [23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 4
- [24] S Vicente and L Agapito. Balloon shapes: Reconstructing and deforming objects with volume from images. In *3DV*, pages 223–230, 2013. 3
- [25] Yufu Wang, Nikos Kolotouros, Kostas Daniilidis, and Marc Badger. Birds of a feather: Capturing avian shape models from images. In *CVPR*, pages 14739–14749, 2021. 3
- [26] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Dove: Learning deformable 3d objects by watching videos. *arXiv preprint arXiv:2107.10844*, 2021. 3
- [27] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *ECCV*, pages 465–481, 2020. 4
- [28] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Learning 3d human shape and pose from dense body parts. In *TPAMI*, 2020. early access. 4
- [29] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019. 4
- [30] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J Black. Three-D safari: Learning to estimate zebra pose, shape, and texture from images “in the wild”. In *ICCV*, pages 5359–5368, 2019. 3
- [31] Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *CVPR*, pages 3955–3963, 2018. 3
- [32] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*, pages 6365–6373, 2017. 2, 3, 7