




BARC: Breed-Augmented Regression Using Classification for 3D Dog Reconstruction from Images

Nadine Rueegg^{1,2}  · Silvia Zuffi³ · Konrad Schindler¹ · Michael J. Black²

Received: 30 April 2022 / Accepted: 1 February 2023 / Published online: 28 April 2023
© The Author(s) 2023

Abstract

The goal of this work is to reconstruct 3D dogs from monocular images. We take a model-based approach, where we estimate the shape and pose parameters of a 3D articulated shape model for dogs. We consider dogs as they constitute a challenging problem, given they are highly articulated and come in a variety of shapes and appearances. Recent work has considered a similar task using the multi-animal SMAL model, with additional limb scale parameters, obtaining reconstructions that are limited in terms of realism. Like previous work, we observe that the original SMAL model is not expressive enough to represent dogs of many different breeds. Moreover, we make the hypothesis that the supervision signal used to train the network, that is 2D keypoints and silhouettes, is not sufficient to learn a regressor that can distinguish between the large variety of dog breeds. We therefore go beyond previous work in two important ways. First, we modify the SMAL shape space to be more appropriate for representing dog shape. Second, 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 is a triplet loss, that encourages the shape of dogs from the same breed to be more similar than dogs of different breeds. The second one is a *breed classification loss*. With our approach we obtain 3D dogs that, compared to previous work, are quantitatively better in terms of 2D reconstruction, and significantly better according to subjective and quantitative 3D evaluations. Our work shows that a-priori side information about similarity of shape and appearance, as provided by breed labels, can help to compensate for the lack of 3D training data. This concept may be applicable to other animal species or groups of species. We call our method BARC (Breed-Augmented Regression using Classification). Our code is publicly available for research purposes at <https://barc.is.tue.mpg.de/>.

Keywords Animal shape estimation · 3D pose estimation · Dogs · Breeds

1 Introduction

The 3D reconstruction of articulated, deformable objects from monocular images is very challenging. Due to the ambiguities in the projection from 3D to 2D, a-priori 3D models of the objects are needed. In the case of humans, recent methods exploit parametric 3D shape models of the human body, like the popular SMPL (Loper et al., 2015), to represent shape. Large collections of 3D body poses, obtained from marker-based motion capture, provide a-priori knowledge about the range of 3D poses (Mahmood et al., 2019). Human body models have been learned from thousands of high-resolution 3D scans, in varied poses, such that pose-dependent deformations can also be encoded. A similar approach cannot be replicated for animals, as scanning them in controlled poses is hard, and in many cases even impossible. Moreover, it may be very difficult to get access to a large number of individuals for

Communicated by Angjoo Kanazawa.

✉ Nadine Rueegg
rueegnad@ethz.ch

Silvia Zuffi
silvia@mi.imati.cnr.it

Konrad Schindler
schindler@ethz.ch

Michael J. Black
black@tue.mpg.de

¹ PRS, ETH Zurich, Zurich, Switzerland

² IS-PS, Max Planck, Tübingen, Germany

³ IMATI, CNR, Milan, Italy



Fig. 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

a specific species of interest. Previous work has learned animal shape using small sets of 3D scans of toy figurines (Zuffi et al., 2017). Even if one learns a parametric shape model of an animal species from a few 3D scans, the limited amount of data will likely restrict the expressive power of that model and it may not be able to capture the shape variability of real individuals. Paired training data of animals, consisting of images and associated, known 3D body shapes, is even rarer. We argue that, to make progress, one must leverage side information that can be obtained more easily, yet constrains 3D shape and pose estimation from single images (Fig. 1).

The 3D reconstruction of animal shape and pose has several applications, ranging from biology and bio-mechanics to entertainment, robotics and conservation. Specifically, the non-invasive capture of 3D body shape supports morphology and health-from-shape analysis, which may be important particularly for endangered species. Markerless motion capture allows 3D motion analysis for animals that are impossible to capture in a lab setting. Animal motion data can be used to drive virtual agents in entertainment, but also to create robots that mimic the often particularly efficient ways in which animals move through their environment. Here we focus on dogs as a rich, challenging test case. Dogs, due to breeding, exhibit an unusually wide range of shapes, and have highly non-rigid, complex articulation. Consequently, they are representative of the variability, and the associated difficulties, of many other animal species.

Our goal is to learn to estimate a dog’s 3D shape and pose from a monocular, uncontrolled image. We consider a supervised, model-based approach, where we train a regressor to predict the parameters of a 3D articulated dog shape model. 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. Moreover, in the absence of 3D ground-truth, unnatural 3D poses can match the 2D keypoints, resulting in bad reconstructions that are not plausible when observed from a different viewpoint. To make the task well-posed, we need additional, prior information. To better estimate 3D *pose*, we define a 3D pose prior based on Normalizing Flows (Kingma et al., 2016; Rezende and Mohamed, 2015). We learn the prior from dog motion capture data, provided in the RGBD-Dog dataset (Kearney et al., 2020). To better estimate dog *shape*, 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 well-suited to explore the role of breed because of their large variety. Dogs have been domesticated and bred for a long time, for diverse purposes such as companionship, hunting, 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 distinct breeds with very different characteristics. A recent analysis of the dog genome illustrates the relationship between different breeds that exist today (Perker et al., 2017). Breeds are grouped into *clades*, often with high shape similarity within a clade. Figure 2 shows a cladogram of 161 domestic dog breeds (Perker et al., 2017). Clades are indicated with colors. Breeds that belong to the same clade are genetically related and therefore share many characteristics, often including shape. A typical example is the European Mastiff clade with the Boxer and the Bulldog. These differ in body size, but have similar build and facial features.

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 neural network called **BARC**, for “Breed-Augmented Regression using Classification.” We

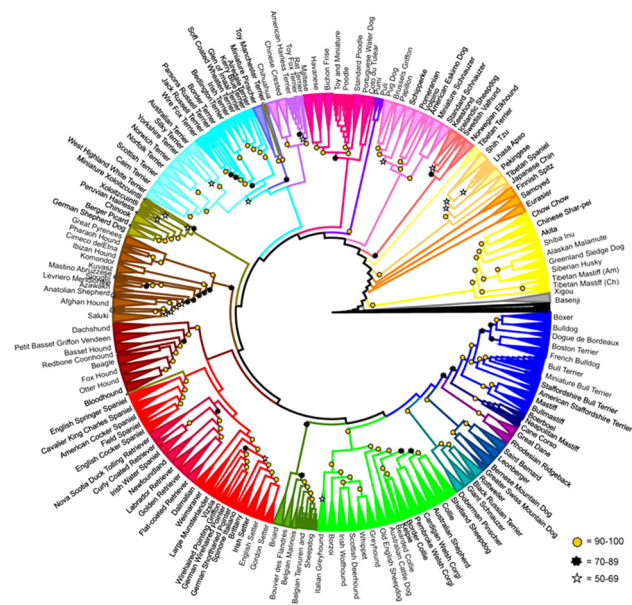


Fig. 2 Cladogram of domestic dog breeds. The diagram represents clustering according to genetic similarity. Reproduced from Perker et al. (2017)

follow the approach of regressing a parametric 3D shape model directly from image pixels, as often done for human pose and shape estimation. Here, we use the SMAL animal model (Zuffi et al., 2017) to define the kinematic chain and mesh template. We extend SMAL in several ways to be a better foundation for learning about dog shape: we add limb scale factors and extend the SMAL shape space with additional 3D examples.

To solve the problem of fitting the 3D model to 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 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 that depend on their breed labels. (3) As a result, we learn a *breed-aware latent shape space*, in which we can identify breed clusters and relationships. An inspection of that shape space, with the help of t-sne plots, indicates good agreement with the cladogram in Fig. 2. (4) Optionally, we show how to exploit 3D models, if they are available for some of the 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 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,

our method is the first attempt to exploit breed information in order to reconstruct 3D animal shape from images.

We train the network on the Stanford Extra (StanExt) (Biggs et al., 2020a; Khosla et al., 2011) training set, which has 120 different breeds. We extend the annotations with eye, withers and throat keypoints. Evaluation is done on the StanExt test set. We find that in the latent shape space that our model learns, closely related dogs are indeed located near each other (Fig. 3). Through ablation studies, we evaluate the impact of different types of breed information and find that each loss term brings a significant improvement in shape accuracy.

To measure accuracy we employ standard 2D measures like PCK and IOU, but these do not fully reflect 3D accuracy. Consequently, we create an additional dataset of 3D dogs, so as to compare the shapes of corresponding breeds. In this way, we are able to carry out a 3D evaluation, in which BARC significantly outperforms the prior art (WLDO (Biggs et al., 2020a)). Finally, to assess shape estimates for in-the-wild images, we carry out a perceptual study, in which we let human subjects compare the 3D dog models visually. We find that BARC reconstructions look more realistic than both those from ablated versions of our model and those from WLDO.

This paper extends an earlier conference publication (Rueegg et al., 2022). Additional content includes details about the body shape model; the 2-dimensional BPS encoding; the body pose prior; and the use of 3D computer graphics models to further constrain body shape. Furthermore, we more comprehensively describe the perceptual evaluation of the reconstructed dog shapes and the evaluation with breed prototypes. We also add an analysis of failure cases, and visual results for ablation experiments where we vary the influence of different loss terms of our model. Finally, we show additional qualitative results, including challenging cases such as puppies and previously unseen breeds.

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.

2.1 Model-Free 3D Reconstruction

These methods do not exploit an existing 3D shape model. Ntouskos et al. (2015) create 3D animal shapes by assembling 3D primitives obtained by fitting manually segmented parts in multiple images of different animals from the same class. Vicente2013 deform a template extracted from a reference image to fit a new image using keypoints and the silhouette,

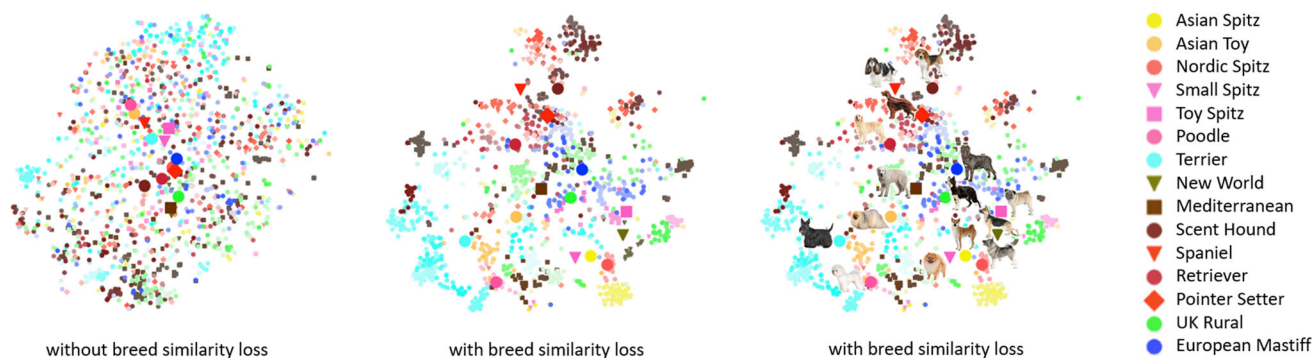


Fig. 3 Learned latent space. t-SNE (Van der Maaten and Hinton, 2008) visualization of the 64-dimensional latent shape variable for dogs in the test set. Large markers indicate average values within each of the clades in Fig. 2. *Left*. Latent space of the network trained without breed simi-

ilarity 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, different saturation levels of the colors denote breeds within the clade (Color figure online)

without addressing articulation. Kanazawa et al. (2018) learn to regress 3D bird shape, given keypoints and silhouettes; birds exhibit rather limited articulation. Recent work obviates the need for 2D keypoints (Goel et al., 2020; Tulsiani et al., 2020; Wu et al., 2021).

2.2 Model-Based 3D Reconstruction

In one of the first 3D animal reconstruction methods from images, Cashman and Fitzgibbon (2013) deform a 3D dolphin template, learning a low-dimensional deformation model from hand-clicked keypoints and a 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. (2017) introduce SMAL, a deformable 3D articulated quadruped animal model. Similar to the widely adopted human body model, SMPL (Loper et al., 2015), 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. (2020a) extend the SMAL model by adding scale parameters for limb lengths. In Wang et al. (2021), 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.

Early work using SMAL employs an optimization-based approach to fit the model to image evidence (Zuffi et al., 2017) and to refine the animal shapes (Zuffi et al., 2018). In other methods, Biggs et al. (2018) show how to extract accurate animal shape and pose from videos, while Kearney et al. (2020) 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. (2020a) 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 shape prior is improved using expectation maximization with respect to fits of their model to the images. Zuffi et al. (2019) 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. (2020) 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.

2.3 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. (2020) 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 (Goel et al., 2020; Kanazawa et al., 2018; Tulsiani et al., 2020) 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

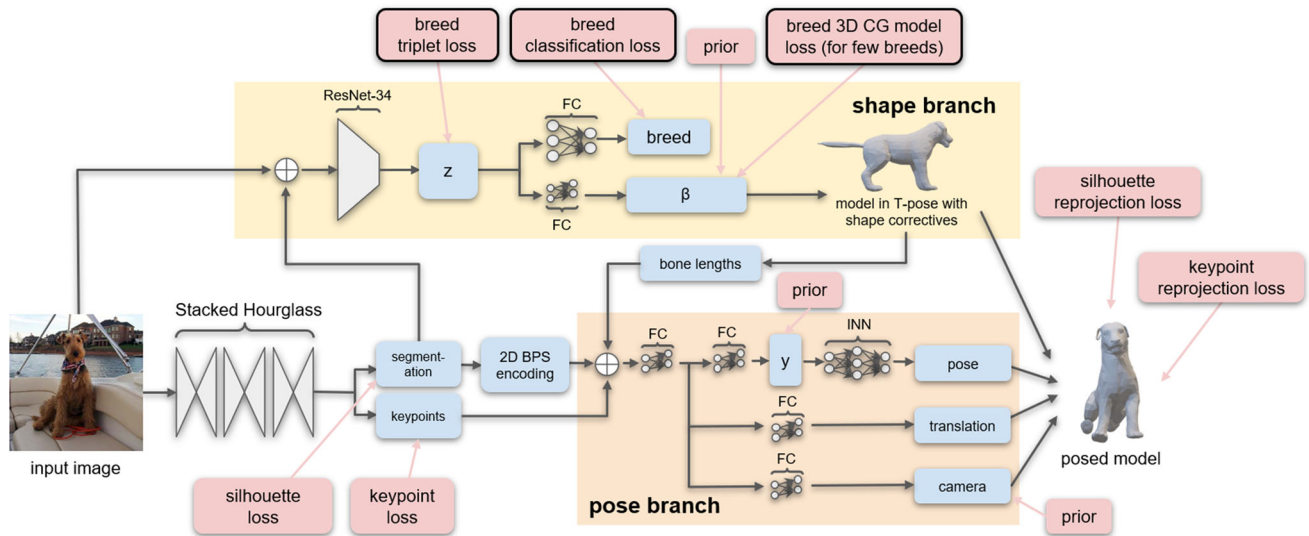


Fig. 4 BARC Architecture. The model consists of a stacked hourglass network followed by two separate branches for shape and pose prediction. Pink boxes illustrate where losses are applied. Black frames indicate the new breed losses

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 backward. We purchased the same pack of Unity models¹ as was exploited by Biggs et al. (2020a) to initialize their mixture 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 κ (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.

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

Strictly speaking, the shape of a specific dog is calculated by following several steps:

1. *Calculating shape deformations caused by β_{pca}* Generic shape directions (PCA directions) are multiplied by instance specific shape coefficients β_{pca} . The result is a vertex wise shape displacement vector.
2. *Adding displacements to shape template:* Those shape displacements are added to a generic shape template (mean shape of PCA). The result is still a mesh of a dog in t-pose.
3. *Posing model and adding shape deformations caused by κ* Linear blend skinning weights (LBS) are used to decide how much influence each limb scaling parameter (exponential) has on each vertex. For example if the exponential of the leg length scaling factor $\kappa_{leg-length}$ is 2, we take each bone of the legs and make it twice as long. Corresponding vertex shifts are calculated based on LBS weights. In practice we do the limb scaling step at the same time as posing the model.

For compactness, we collect the PCA shape coefficients β_{pca} and limb scales κ into a shape vector β .

3.2 Architecture

Similar to Pavlakos et al. (2018) and Zhang et al. (2020), 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.

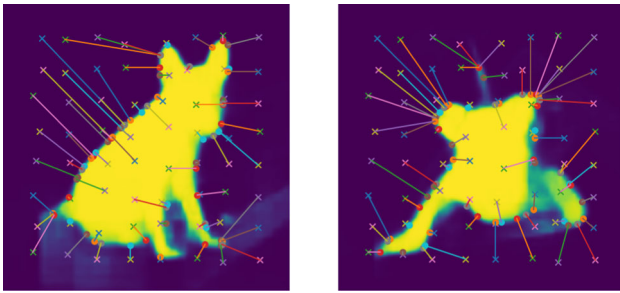


Fig. 5 Silhouette BPS encoding. Crosses indicate basis points and corresponding dots the approximately closest points on the silhouette circumference

3.2.1 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, Nibali et al., 2018). The segmentation map is encoded with a scheme similar to “basis point sets” (BPS, Prokudin et al., 2019) for 3D point cloud encoding. To our knowledge, we are the first to apply BPS in 2D.

Figure 5 illustrates 64 points that form our basis point set (crosses) and the corresponding closest points on the silhouette (dots). The basis points are predefined, by overlaying a regular 8×8 grid on the image and adding a small random perturbation independently to each grid point. We concatenate the x - and y -coordinates of all silhouette points to obtain a 128-dimensional encoding. 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 improves the 3D prediction over 2D keypoints alone. Notably, this encoding is efficient in terms of model size and does not require any training, in contrast to alternatives such as convolutional silhouette features.

3.2.2 Shape-Branch

The input image and the predicted segmentation map are concatenated and fed to a ResNet34 (He et al., 2016) that predicts 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 β . We have experimented with different sub-networks between z and β 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 κ and β_{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.

3.2.3 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 with respect to the camera coordinate system and the camera’s focal length. The pose is represented as a 6D rotation (Zhou et al., 2019) 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 (Xu et al., 2020; Biggs et al., 2020b; Zanfir et al., 2020), 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 (Kearney et al., 2020). Similar to Zanfir et al. (2020), this network consists of Real-NVP blocks, but because of the smaller size of the RGBD-Dog dataset [compared to AMASS (Mahmood et al., 2019)] our pose network is much smaller than those previously used for human pose estimation. For architectural details, please refer to the code, available online. 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 (Kearney et al., 2020), 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.

3.2.4 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 (Ravi et al., 2020).

3.3 Training Procedure

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

3.3.1 Stacked Hourglass Pretraining

The stacked hourglass is pretrained to predict keypoints and the segmentation map. The StanExt dog dataset (Biggs et al., 2020a) 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.

3.3.2 Pose-Branch Pretraining

We use the same dataset (RGBD-Dog) that is used to train the pose prior to also pretrain the pose branch. We sample poses and random shapes and project them to a 256×256 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.

3.3.3 Main Training

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

3.4 Standard Losses

3.4.1 Keypoint Reprojection Loss

L^{kp} is the weighted mean squared error between predicted k_n^{pred} and ground truth 2D keypoint locations k_n^{gt} :

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

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 are listed in Table 1 and balance the influence of keypoints.

3.4.2 Silhouette Reprojection Loss

L^{sil} is the squared pixel error between the rendered s^{pred} and ground truth silhouette s^{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 .

Table 1 Keypoint weights

Keypoint	w	Keypoint	w
Left front leg, paw	3	Right rear leg, top	2
Left front leg, middle	2	Tail start	3
Left front leg, top	2	Tail end	3
Left rear leg, paw	3	Base left ear	2
Left rear leg, middle	2	Base right ear	2
Left rear leg, top	2	Nose	3
Right front leg, paw	3	Chin	1
Right front leg, middle	2	Left ear tip	2
Right front leg, top	2	Right ear tip	2
Right rear leg, paw	3	Left eye	1
Right rear leg, middle	2	Right eye	1

Weights that are used within the weighted keypoint loss

3.4.3 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 μ_{pca} and covariance Σ_{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 κ ,

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

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

3.4.4 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.

We calculate this directly from our latent representation y under a multivariate normal distribution with mean vector μ and covariance matrix Σ . We can write the multivariate normal probability density function $f(y)$ evaluated at a vector y of dimension d using the Mahalanobis distance between this vector and μ , $MD(y; \mu, \Sigma) = (y - \mu)^T \Sigma^{-1} (y - \mu)$ as:

$$f(y) = \frac{1}{\sqrt{(2\pi)^d \|\Sigma\|}} \exp\left(-\frac{1}{2} MD(y; \mu, \Sigma)\right), \quad (5)$$

and thus the log likelihood, $\log(f(y))$

$$= -\frac{1}{2}(d \log(2\pi) + \log(\|\Sigma\|) + MD(y; \mu, \Sigma)^2). \tag{6}$$

Since our normalizing flow prior is trained under the assumption of zero mean and unit variance this can be simplified to obtain:

$$\log(f(y)) = -\frac{1}{2}(d \log(2\pi) + y^T y). \tag{7}$$

Finally, the normalizing flow pose loss is given by:

$$L_{nf}^p = \frac{1}{2}(d \log(2\pi) + y^T y). \tag{8}$$

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_{side}^p that penalizes sideways poses of the joints in each leg. The idea behind this term is that leg joints in the SMAL model are socket joints, but in reality they are hinge joints. Therefore, by penalizing rotations to the side, we penalize rotations that would be anatomically incorrect. The final pose prior is:

$$L^p = w_{nf}L_{nf}^p + w_{side}L_{side}^p, \tag{9}$$

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

3.4.5 Camera Prior

L^{cam} : Since focal length f^{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^{target} :

$$L^{cam} = (f^{pred} - f^{target})^2. \tag{10}$$

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.

3.5.1 Breed Triplet Loss

$L_{triplet}^B$: 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 (Schroff et al., 2015; Taigman et al., 2014). RingNet used a similar idea to learn 3D head shape from images without 3D supervision (Sanyal et al., 2019).

Applying the loss directly to the shape β 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 β_{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, $L_{triplet}^B =$

$$\sum_{i=1}^{N_{triplets}} \max(d(z_{a,i}, z_{p,i}) - d(z_{n,i}, z_{a,i}) + m, 0), \tag{11}$$

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

3.5.2 Breed Classification Loss

L_{cs}^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:

$$L_{cs}^B = - \sum_{c=1}^{N_{classes}} y_{o,c} \log(p_{o,c}), \tag{12}$$

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:

$$L_{sim}^B = w_{triplet}L_{triplet}^B + w_{cs}L_{cs}^B, \tag{13}$$

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

3.5.3 3D Model Loss

L_{3D}^B : We have access to a small number of 3D dogs (*Unity* models) and to a few 3D scans of toy figurines, namely the canine examples of the SMAL training set, which we also used to construct the dog model, see Sec. 3.1. 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 β :

$$L_{3D}^B = (\beta_{pca}^{pred} - \beta_{pca}^{breed})^2 + (\kappa^{pred} - \kappa^{breed})^2. \quad (14)$$

Table 2 shows a list of 3D CG models and corresponding breeds which BARC uses in its 3D model loss.

4 Experiments

We evaluate our approach on the Stanford Extra Dog dataset (StanExt) (Biggs et al., 2020a). 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. Those keypoints are obtained by training a separate stacked hourglass on the Animal Pose dataset (Cao et al., 2019) and using its predictions as pseudo ground truth in the StanExt training set.

4.1 Evaluation Methods

4.1.1 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).

4.1.2 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 more 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.

Controlled perceptual tasks are designed to evaluate our method relative to (1) the SOTA or (2) to an ablated model. Workers on Amazon Mechanical Turk (AMT) judge which of two rendered 3D body shapes better fits a query dog image. Figure 6 shows the framework that we provide to the AMT workers. We show each worker an image that contains a dog, our predicted 3D model in T-pose and the model in T-pose from SOTA or ablated method. We do not present the predicted 3D posed models in order to focus workers on shape. The left-right ordering of the rendered meshes is random. We let each worker first process 8 samples to get used to the task, and then use the next 30 hits. The task is split in 4 batches with 30 samples each. We have 10 workers for each batch. This gives us a total of 1200 hits. In order to verify the workers understand the task and perform it diligently, we include two catch trials in each batch. These are extreme cases where



Fig. 6 AMT Framework. The picture shows an example screenshot from the perceptual studies that we ran on Amazon Mechanical Turk

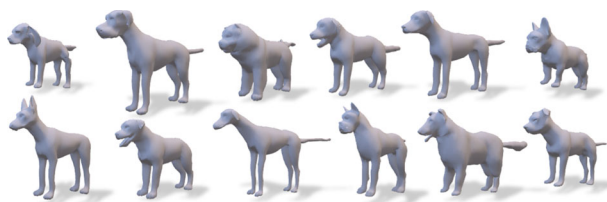


Fig. 7 Breed Prototypes. From top left: Beagle, Great Dane, Chow-Chow, Labrador Retriever, Rhodesian Ridgeback, French Bulldog, Doberman Pinscher, Rottweiler, Greyhound, Boxer, Collie, American Staffordshire Terrier

one 3D shape is so far off that only one answer is plausible. For all quantitative results reported, votes from workers who failed one or both catch trials are ignored.

4.1.3 Breed Prototype Consistency

We complement the perceptual shape evaluation by an evaluation which exploits the fact that dogs of the same breed 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. We use 20 prototypes. Figure 7 shows a few examples. 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 between the estimate and the prototype serves as indicator of how well a given prediction method captures the breed shape.

Table 2 3D CG models

Breed	Stanford extra name
American Staffordshire Terrier	n02093428-American_Staffordshire_terrier
Boxer	n02108089-boxer
German Shepherd	n02106662-German_shepherd
Doberman	n02107142-Doberman
Staffordshire Bullterrier	n02093256-Staffordshire_bullterrier
French Bulldog	n02108915-French_bulldog
Bull Mastiff	n02108422-bull_mastiff
Great Dane	n02109047-Great_Dane
Italian Greyhound	n02091032-Italian_greyhound
Rottweiler	n02106550-Rottweiler
Siberian Husky	n02110185-Siberian_husky

Models used for our 3D model loss L_{3D}^B

Table 3 Comparison to SOTA

Method	IoU	PCK @ 0.15				
		Avg	Legs	Tail	Ears	Face
3D-M	69.9	69.7	68.3	68.0	57.8	93.7
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
Ours	75.1	82.8	82.3	63.3	83.3	91.3

Bold means better

Numbers for 3D-M (Zuffi et al., 2017), CGAS (Biggs et al., 2018), WLDO (Biggs et al., 2020a) reproduced from Biggs et al. (2020a)

4.2 Comparison to Baselines

In terms of 2D error metrics (IoU and PCK) BARC outperforms prior art, i.e., WLDO (Biggs et al., 2020a), CGAS (Biggs et al., 2018) and 3D-M (Zuffi et al., 2017). Table 3 summarizes the results. Importantly, the methods compared in the table exploit different types and amounts of information during training. In particular, our method has access to breed labels, which WLDO has not not. Moreover, our method employs additional keypoints for the eyes and the throat, and a quite different pose prior. Also in the perceptual comparison, BARC is judged to represent the depicted dog better than its closest competitor WLDO, in an overwhelming 92.4% of all cases. See last line of Table 4. The marked gap in visual realism is evident in Fig. 8.

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 (Table 4) and in terms of consistency with the prototype breed shape (Table 5). We compare the following versions of our



Fig. 8 Comparison to SOTA. Qualitative comparison of BARC (left half) with WLDO (Biggs et al., 2020a) (right half). For each method we show the input image, the 3D reconstruction projected on the input image, the 3D reconstruction, and a 90° rotated view

method: (i) our network, trained without any breed losses; (ii-a) the same network with classification loss L_{cs}^B ; (ii-b) with full breed similarity losses L_{sim}^B , i.e., classification L_{cs}^B and triplet $L_{triplet}^B$ loss; (iii) with all breed losses, including the 3D CG model loss L_{3D}^B .

4.3.1 Perceptual Shape Evaluation

Table 4 shows that in terms of perceptual agreement, the two parts, similarity loss and 3D CG model loss, have similar impact.

Even though they do not explicitly constrain the 3D shape, triplet and classification loss bring a clear improvement.

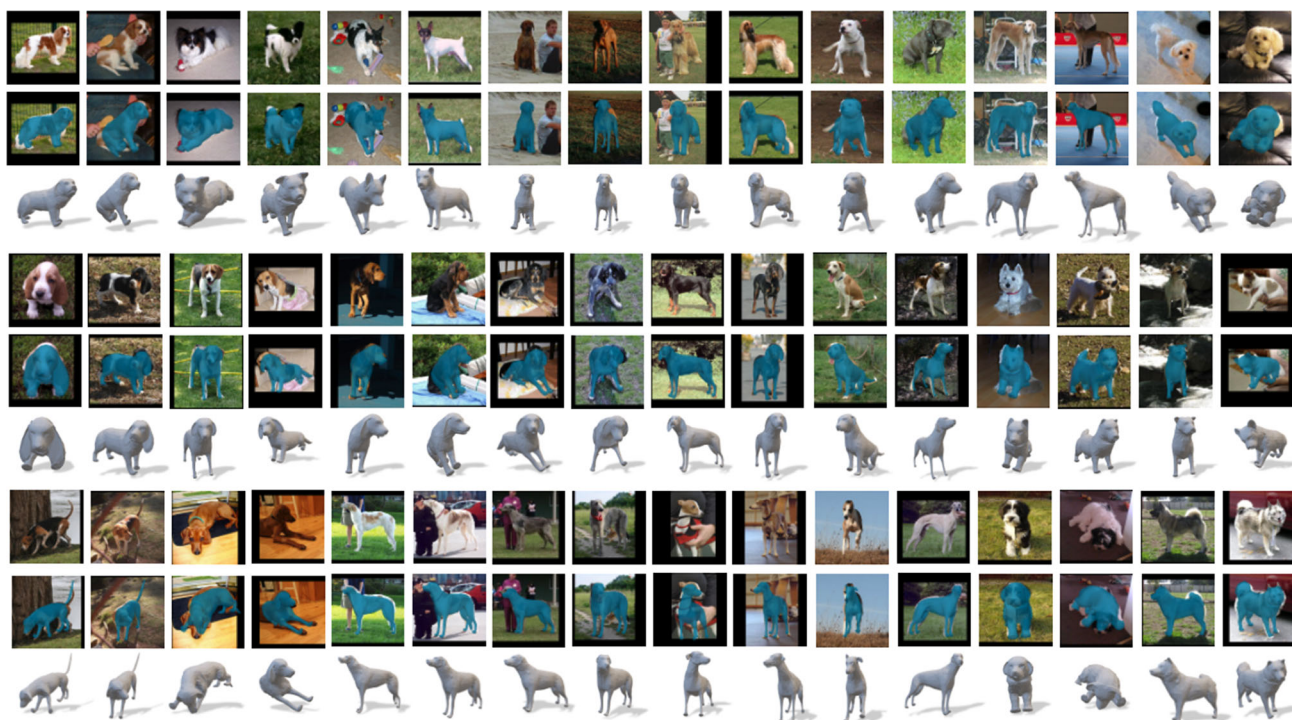


Fig. 9 BARC results. Each row shows the input image with the projected 3D shape. Below that is an overlay of our predicted model and the image, and finally a rendering of the posed 3D shape

Table 4 Perceptual studies

Experiment settings	AMT results	
	Votes	Percentage
L_{sim}^B versus no breed losses	638: 382	62.55% : 37.45%
$\{L_{sim}^B, L_{3D}^B\}$ versus L_{sim}^B	670: 440	60.36% : 39.64%
$\{L_{sim}^B, L_{3D}^B\}$ versus WLDO	998: 82	92.41% : 7.59%

Bold means better

Ablation of breed losses and comparison with WLDO. See text

Breed-specific 3D shape information as exploit by the 3D CG model loss 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 (χ^2 -test, $p < 0.0001$).

4.3.2 Breed Prototype Consistency

We complement the perceptual study with a quantitative evaluation with respect to breed prototypes (Table 5).

For 20 different breeds we evaluate WLDO, as well as our method without any breed losses, with only classification loss L_{cs}^B , with classification and triples loss summarized as similarity loss L_{sim}^B , and with both L_{sim}^B and L_{3D}^B . Already

Table 5 3D shape evaluation

Method	WLDO	BARC			
		None	L_{cs}^B	L_{sim}^B	$\{L_{sim}^B, L_{3D}^B\}$
Error (cm)	11.55	8.58	7.99	7.76	6.95

Bold means better

For different breed losses, breed prototype consistency averaged over 20 breeds

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 classification and triplet losses decreases the error further. The additional 3D breed loss brings another reduction, 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, $WLDO > BARC_{nobreed} > BARC_{sim} > BARC_{sim+3D}$.

4.3.3 T-SNE Visualizations

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)

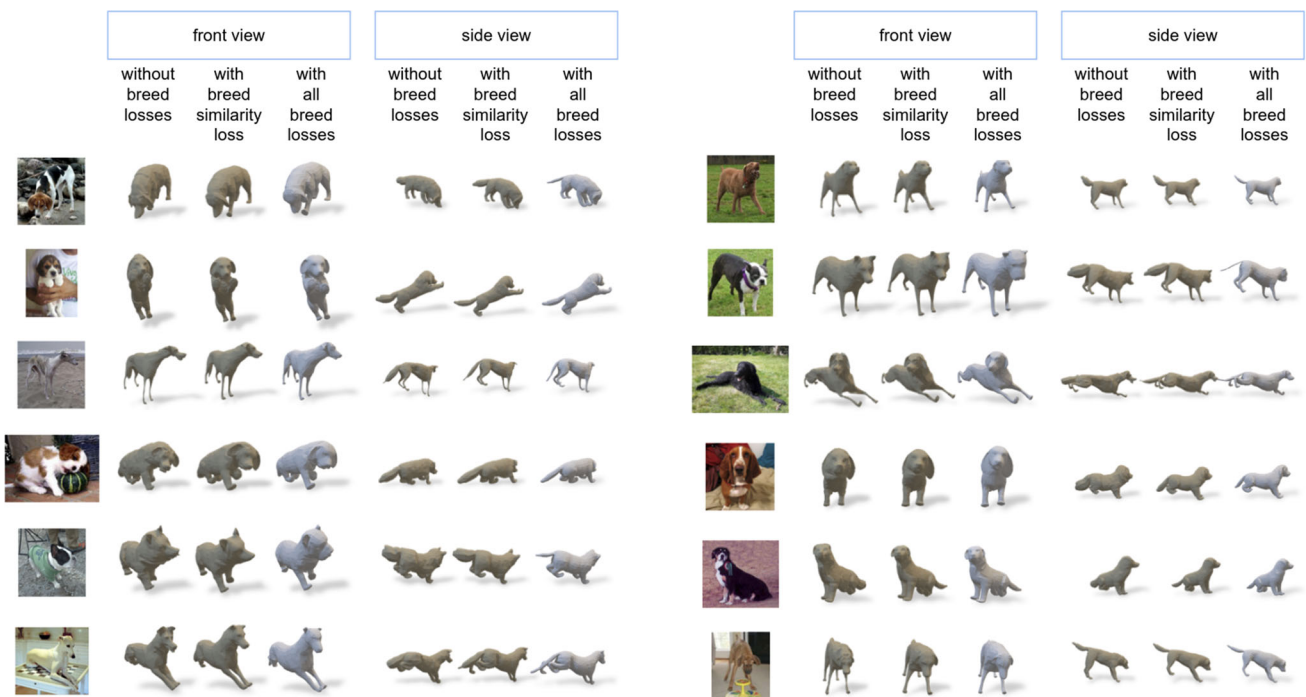


Fig. 10 Ablation study. Qualitative comparison of from left to right (1) our method trained without any breed losses (2) our method trained with similarity breed loss only (3) BARC (our method). We show for various input images, front views as well as side views

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 β . Different saturation levels of the same color indicate breeds within the clade and clade colors are kept consistent with the colors of the cladogram in Fig. 2. 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 (Fig. 9).

4.3.4 Qualitative Results for Ablated Models

Figure 10 shows results for ablated versions of BARC. To the left we render results from our method without any of the breed related losses, in the middle results with the breed similarity loss only and to the right with the breed similarity loss as well as the 3D CG model loss. For each of the three versions, we show the front as well as a side view.

4.4 Failure Case Analysis

We divide the failure cases in two main groups: shape and pose failures.



Fig. 11 Failure cases. Most failure cases are due to occlusion and poses not seen during training

4.4.1 Pose Failure Cases

At development time we have trained our network with various pose priors, such as a mixture of Gaussians prior as in Zuffi et al. (2018) and Biggs et al. (2020a), a variational auto-encoder as in Zuffi et al. (2019) and our final normalizing flow pose prior. One failure mode that goes through all priors is the erroneous prediction of dogs not facing the camera. The Stanford Extra training set is unbalanced in the

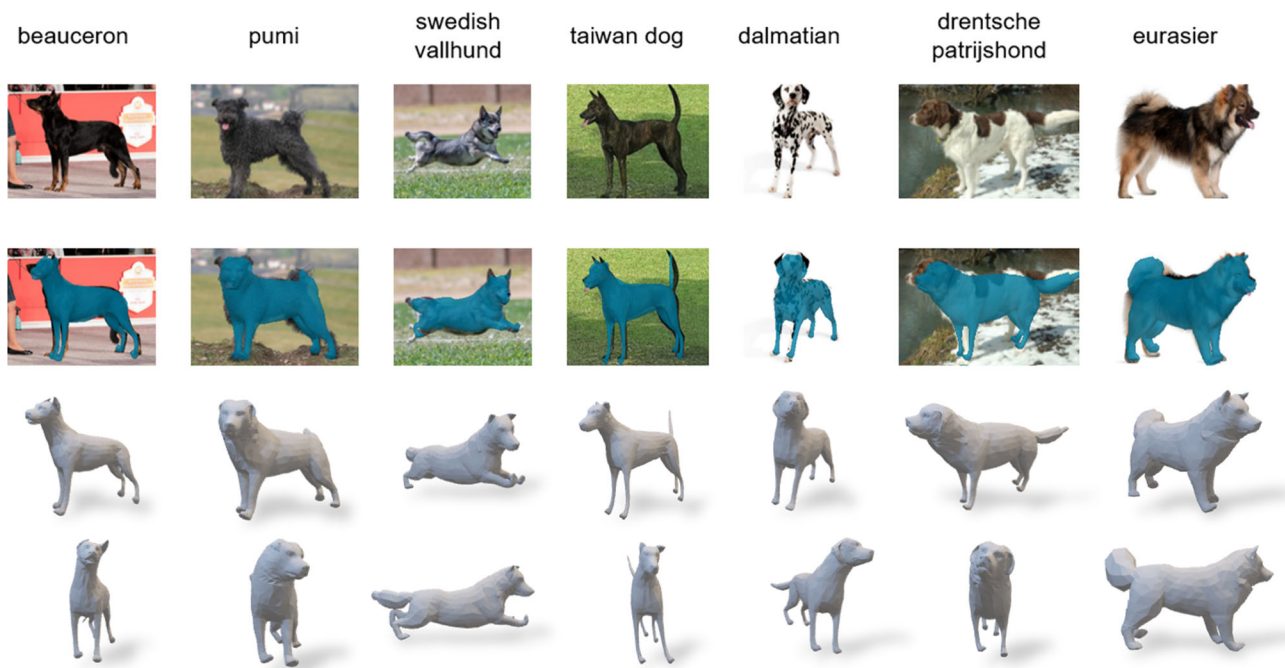


Fig. 12 Results for unseen breeds. Qualitative results of BARC (our method) on images of previously unseen breeds. All test images are downloaded from the American Kennel Club web page. We show for various input images an overlay, front view as well as side view of our predicted dog

sense that it shows many dogs from a front- or side-view. Furthermore, most of the dogs do not bend the front legs as they are either sitting, laying or standing, this leads to challenges when predicting poses for dogs with heavily bent wrists. As training with different pose priors lead to similar error cases, we believe that those challenges are not structural problems of the pose prior itself, but rather of the image dataset. Nevertheless, it might be worth examining different training schedules such that rare poses obtain higher weights or are repeated more often. One more thing worth mentioning is that often perceived 3D quality from front view is considerably higher than from side-views. A strong 3D regularization is inevitable. Predictions for laying and sitting dogs could be improved by training a pose prior on a more suitable 3D pose dataset. Furthermore, BARC is not always able to correctly predict the pose if the dog is only partly visible, or if its pose is far from those seen at training time. All methods, including BARC, fail completely for a few images and predict translations where the dog does not even project into the image.

4.4.2 Shape Failure Cases

Our breed losses help to regularize dog shape. BARC can predict more reasonable shapes, especially for dogs that are not fully visible from the side. Nevertheless, we do sometimes observe shortened limbs when they are difficult to predict due to poses such as a dog laying and facing the camera. Working with a single shape for each dog breed is



Fig. 13 Puppies. Qualitative results on puppies from the Stanford Extra test set

not an option, as there is no negligible intra-class variability. Another challenge is dog hair. First, shape variability can become enormous, consider for example differently sheared poodles. Secondly, long hair does swing and the shape that we want to predict for a dog with fluffy hair is not clearly defined. In such cases, representing a dog with a mesh is not ideal.

4.4.3 Some Visual Examples of Failure Cases

We show four failure cases in Fig. 11:

- (1) a dog which is not fully visible, our prediction shows a shrunken body.
- (2) most training images show dogs that face

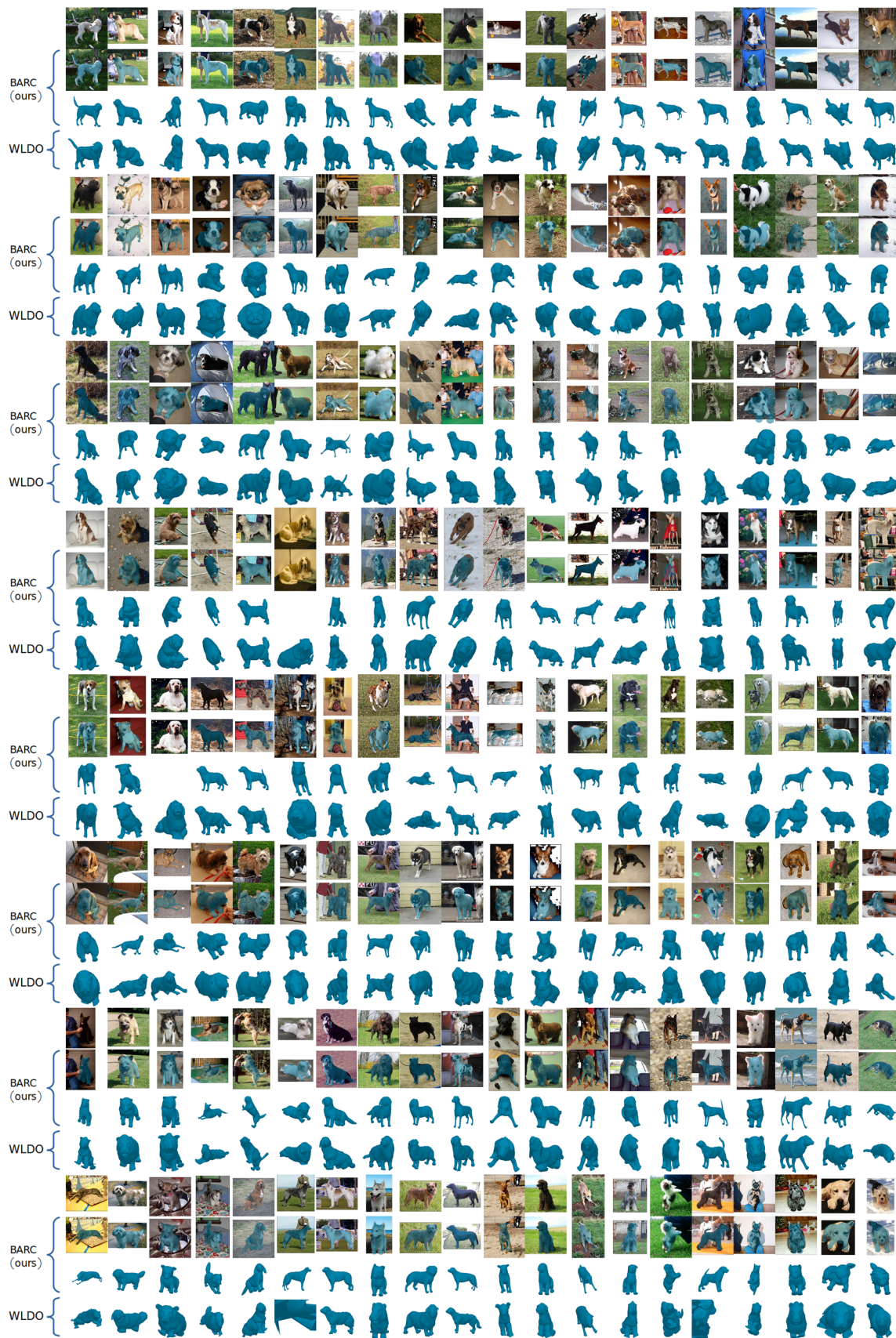


Fig. 14 Randomly sampled results. We show qualitative results on the Stanford Extra test set: for each sample an input image, the overlay of our prediction (BARC) with that image, our prediction and previous state-of-the-art (WLDO)

the camera. When the dog is turned away, pose prediction fails. (3) a Japanese Spaniel with lots of hair. Shape prediction for such breeds is difficult. (4) A dog that is hard to recognize and where, in part, the difficult pose is compensated by a wrong shape - instead of bending the back, the dog is given a stouter body.

4.5 BARC Visualized

In this section we present qualitative results of our final method BARC and show that BARC generalizes well to new dog shapes. BARC results on StanExt test images for different breeds, are displayed in Fig. 9. While input images in this figure are sampled for variety, we use Fig. 14 to present results on completely randomly sampled Stanford Extra test set images. For each input image we show the overlap of our prediction with this image, a 3D visualization of our prediction and a 3D visualization of the previous state-of-the-art method WLDO.

Finally, we aim to gain insights with respect to the generalization ability of BARC to new shapes. StanExt contains images of dogs belonging to 120 different breeds, and we are interested to see how BARC performs on previously unseen breeds. To that aim, images of new dog breeds are downloaded from the American Kennel Club web page. Figure 12 illustrates an overlay of our prediction on the input image, as well as front and side view for each of the seven dogs. In a second experiment, we direct our attention to puppies. While a few puppies are part of the StanExt training set, they are not labelled as such and treated similarly as the adult dogs. We visualize results on puppy images in Fig. 13. We provide for each example, the input image, an overlay of the prediction with that image, a rendering of the posed 3D model and a rendering of the dog's shape in t-pose. We conclude that BARC generalizes to previously unseen breeds as well as puppies (Fig. 14).

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. 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.

5.1 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, and veterinary medicine.

Acknowledgements This research was supported by the Max Planck ETH Center for Learning Systems. MJB has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. MJB has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH. While MJB was a part-time employee of Amazon during this Project, his research was performed solely at, and funded solely by the Max Planck Society.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Biggs, B., Boyne, O., Charles, J., Fitzgibbon, A., & Cipolla, R. (2020a). Who left the dogs out: 3D animal reconstruction with expectation maximization in the loop. In *ECCV*.
- Biggs, B., Ehrhardt, S., & Joo, H., Graham, B., & Vedaldi, A. 3D multi-bodies: Fitting sets of plausible 3D human models to ambiguous image data. arXiv preprint [arXiv:2011.00980](https://arxiv.org/abs/2011.00980)
- Biggs, B., Roddick, T., & Fitzgibbon, A., & Cipolla, R. (2018). Creatures great and SMAL: Recovering the shape and motion of animals from video. In *ACCV*.
- Cao, J., Tang, H., Fang, H.S., Shen, X., Lu, C., & Tai, Y. W. (2019). Cross-domain adaptation for animal pose estimation. In *ICCV*.
- Cashman, T. J., & Fitzgibbon, A. W. (2013). What shape are dolphins? Building 3D morphable models from 2D images. *IEEE TPAMI*, 35(1), 232–244.

- Goel, S., Kanazawa, A., & Malik, J. (2020). Shape and viewpoints without keypoints. In *ECCV*.
- He, K., Zhang, X., Ren, S., & Sun J. (2016). Deep residual learning for image recognition. In *CVPR*.
- Kanazawa, A., Tulsiani, S., Efros, A. A., & Malik, J. (2018). Learning category-specific mesh reconstruction from image collections. In *ECCV*.
- Kearney, S., Li, W., & Parsons, M., Kim, K. I., & Cosker, D. (2020). RGBD-dog: Predicting canine pose from RGBD sensors. In *CVPR*
- Khosla, A., Jayadevaprakash, N., Yao, B., & Li, F. F. (2011). Novel dataset for fine-grained image categorization. In *CVPR workshops*.
- Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., & Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *NIPS*.
- Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM ToG—SIGGRAPH Asia*, 34(6), 1–16.
- Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., & Black, M. J. (2019). AMASS: Archive of motion capture as surface shapes. In *ICCV*.
- Mu, J., Qiu, W., Hager, G. D., & Yuille, A. L. (2020). Learning from synthetic animals. In *CVPR*
- Nibali, A., He, Z., Morgan, S., & Prendergast, L. (2018). Numerical coordinate regression with convolutional neural networks. arXiv preprint [arXiv:1801.07372](https://arxiv.org/abs/1801.07372)
- Ntoutoskos, V., Sanzari, M., Cafaro, B., Nardi, F., Natola, F., Pirri, F., & Ruiz, M. (2015). Component-wise modeling of articulated objects. In *ICCV*.
- Parker, H. G., Dreger, D. L., Rimbault, M., Davis, B. W., Mullen, A. B., Carpintero-Ramirez, G., & Ostrander, E. A. (2017). Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. *Cell Report*, 4(19), 697–708.
- Pavlakos, G., Zhu, L., Zhou, X., & Daniilidis, K. (2018). Learning to estimate 3D human pose and shape from a single color image. In *CVPR*.
- Prokudin, S., Lassner, C., & Romero, J. (2019). Efficient learning on point clouds with basis point sets. In *ICCV*
- Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., & Gkioxari, G. (2020). Accelerating 3D deep learning with PyTorch3D. [arXiv:2007.08501](https://arxiv.org/abs/2007.08501)
- Rezende, D., & Mohamed, S. (2015). Variational inference with normalizing flows. In *ICML*.
- Rueegg, N., Zuffi, S., Schindler, K., & Black, M. J. (2022). Barc: Learning to regress 3d dog shape from images by exploiting breed information. In *CVPR*.
- Sanakoyeu, A., Khalidov, V., McCarthy, M. S., Vedaldi, A., & Neverova, N. (2020) Transferring dense pose to proximal animal classes. In *CVPR*.
- Sanyal, S., Bolkart, T., Feng, H., & Black, M. J. (2019). Learning to regress 3D face shape and expression from an image without 3D supervision. In *CVPR*.
- Schroff, F., Kalenichenko, D., & Philbin, J. (2015) FaceNet: A unified embedding for face recognition and clustering. In *CVPR*.
- Taijman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. In *CVPR*.
- Tulsiani, S., Kulkarni, N., & Gupta, A. (2020). Implicit mesh reconstruction from unannotated image collections. arXiv preprint [arXiv:2007.08504](https://arxiv.org/abs/2007.08504)
- Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11).
- Vicente, S., & Agapito, L. (2013). Balloon shapes: Reconstructing and deforming objects with volume from images. In *3DV*.
- Wang, Y., Kolotouros, N., Daniilidis, K., & Badger, M. (2021). Birds of a feather: Capturing avian shape models from images. In *CVPR*.
- Wu, S., Jakob, T., Rupprecht, C., & Vedaldi, A. (2021). Dove: Learning deformable 3D objects by watching videos. arXiv preprint [arXiv:2107.10844](https://arxiv.org/abs/2107.10844)
- Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W. T., Sukthankar, R., & Sminchisescu, C. (2020). Ghum & ghum!: Generative 3D human shape and articulated pose models. In *CVPR*.
- Zanfir, A., Bazavan, E.G., Xu, H., Freeman, W.T., Sukthankar, R., & Sminchisescu, C. (2020). Weakly supervised 3D human pose and shape reconstruction with normalizing flows. In *ECCV*.
- Zhang, H., Cao, J., Lu, G., Ouyang, W., & Sun, Z. (2020). Learning 3D human shape and pose from dense body parts. In *TPAMI*.
- Zhou, Y., Barnes, C., Lu, J., Yang, J., & Li, H. (2019) On the continuity of rotation representations in neural networks. In *CVPR*.
- Zuffi, S., Kanazawa, A., Berger-Wolf, T., & Black, M. J. (2019). Three-D safari: Learning to estimate zebra pose, shape, and texture from images “in the wild”. In *ICCV*.
- Zuffi, S., Kanazawa, A., & Black, M. J. (2018). Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *CVPR*.
- Zuffi, S., Kanazawa, A., Jacobs, D. W., & Black, M. J. (2017) 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.