BITE: Beyond Priors for Improved Three-D Dog Pose Estimation

Nadine Rüegg[1,2], Shashank Tripathi[2], Konrad Schindler[1], Michael J. Black[2], and Silvia Zuffi[3]

[1]ETH Zürich, Switzerland
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3]IMATI-CNR, Milan, Italy

Figure 1. BITE enables 3D shape and pose estimation of dogs from a single input image. The model handles a wide range of shapes and breeds, as well as challenging postures far from the available training poses, like sitting or lying on the ground.

Abstract

We address the problem of inferring the 3D shape and pose of dogs from images. Given the lack of 3D training data, this problem is challenging, and the best methods lag behind those designed to estimate human shape and pose. To make progress, we attack the problem from multiple sides at once. First, we need a good 3D shape prior, like those available for humans. To that end, we learn a dog-specific 3D parametric model, called D-SMAL. Second, existing methods focus on dogs in standing poses because when they sit or lie down, their legs are self occluded and their bodies deform. Without access to a good pose prior or 3D data, we need an alternative approach. To that end, we exploit contact with the ground as a form of side information. We consider an existing large dataset of dog images and label any 3D contact of the dog with the ground. We exploit body-ground contact in estimating dog pose and find that it significantly improves results. Third, we develop a novel neural network architecture to infer and exploit this contact information. Fourth, to make progress, we have to be able to measure it. Current evaluation metrics are based on 2D features like keypoints and silhouettes, which do not directly correlate with 3D errors. To address this, we create a synthetic dataset containing rendered images of scanned 3D dogs. With these advances, our method recovers significantly better dog shape and pose than the state of the art, and we evaluate this improvement in 3D. Our code, model and test dataset are publicly available for research purposes at https://bite.is.tue.mpg.de/.

1. Introduction

Capturing and modeling 3D animal shape and pose has many applications, ranging from biology and conservation to entertainment and virtual content creation. Cameras are a natural sensor to observe animals because they do not require the animal to stand still, hold specific postures, be in physical contact, or otherwise cooperate. The study of animals using images has a long tradition, e.g. Muybridge's famous "Horse in Motion" chronophotographs [30]. Still, expressive 3D models that can adapt to the individual shape and pose of an animal have only recently been created [44], following earlier work on 3D human shape and pose [1, 23, 38]. Here, we address the task of reconstructing dogs in 3D from a single image. We focus on dogs as a representative animal species that features both large shape variability across different breeds and strong articulated deformations typical of quadrupeds. Dogs are frequently photographed, so images showing a wide range of poses, shapes and environments are readily available.

While, at first glance, modeling humans and modeling dogs may seem like similar problems, they present very different technical challenges. For humans there is an enor-

mous amount of existing 3D scan and motion capture data. This data covers relevant pose and shape variations, which has made it possible to learn powerful, articulated models like SMPL [23] or GHUM [38].[1] In contrast, 3D observations of animals are difficult to acquire, and at present too scarce to train 3D statistical models that are equally expressive and cover all possible shapes and poses. The introduction of SMAL [44], a parametric quadruped model learned from toy figurines, has enabled significant progress and has made it possible to reconstruct animals in 3D from images, including dogs [5, 21, 26]. However, SMAL is a generic model for multiple species from cats to hippos. While it can represent the varied body shapes of different animals, it cannot represent the distinctive and fine-grained characteristics of different dog breeds (e.g., the wide variety of ears). To address this issue, we introduce the first dog-specific parametric model, termed D-SMAL, to accurately represent dogs.

An additional issue is that there is very little motion capture data of dogs (unlike humans) and the data that does exist rarely captures sitting and lying poses. This makes it difficult for existing methods to infer dogs in such poses. For example, if one learns a prior over 3D poses from existing data, it will be biased to standing/walking poses. One could weaken this prior, using generic constraints, but then the estimation of pose is highly under-constrained. To address this problem, we exploit information about physical contact that has been neglected when modeling (land) animals: they are subject to gravity and therefore stand, sit or lie on the ground. We introduce ground contact information and show that we can exploit this to estimate complex dog poses, even in challenging cases with significant self-occlusion. While ground plane constraints have been exploited in human pose estimation, their potential benefit is larger for quadrupeds, simply because four legs mean more ground contact points, as well as more occluded body parts and larger non-rigid deformations when sitting or lying down.

Another limitation of previous work is that the reconstruction pipelines are usually trained on 2D images, due to the difficulty of collecting 3D data (with paired 2D images). As a result, they tend to predict shapes and poses that accurately match the image evidence when re-projected, but are distorted along the viewing direction. Viewed from a different angle, the 3D reconstruction may be inaccurate because, in the absence of paired data, there is not enough evidence to learn where to place more distant or even occluded body parts along the depth direction. Again, we find that modelling ground contact helps: to circumvent the manual reconstruction (or synthesis) necessary to obtain paired 2D and 3D data, we revert to a weaker form of 3D supervision and obtain ground contact labels. Specifically, we present real images to annotators and ask them to label whether the ground surface under the dog is flat and, if so, to also annotate the ground contact points on the 3D animal. These labels are exploited not only for training: we found that the network can be trained to classify the surface and detect the contact points fairly reliably from a single image, such that they can also be used at test time.

We call our reconstruction system BITE, which we base on the current state-of-the-art model, BARC [26]. As an initial, coarse fitting stage we re-train BARC with our new D-SMAL dog model. We then pass the resulting predictions to our newly developed refinement network, which we train with ground contact losses to refine the dog's pose (as well as the camera parameters). Optionally, at test time, we can further exploit the ground contact loss to optimize the fit to the test image, fully automatically. This significantly improves reconstruction quality. With BITE, we obtain dogs that correctly stand on the (locally planar) ground, or are reconstructed realistically in sitting and lying postures, even though the training set for the BARC pose prior does not contain such poses (see Fig. 1).

Previous work on 3D dog reconstruction is evaluated by back-projecting to the image and measuring 2D residuals, thus projecting away errors in depth [5], or via subjective visual ratings [26]. To address the lack of objective 3D evaluations, we have created a novel, semi-synthetic dataset with 3D ground truth, by rendering 3D scans of real dogs from different viewing angles. We evaluate BITE as well as its main competitors on this new dataset, and show that BITE indeed sets a new state of the art.

In summary, our contributions are:

1. We introduce D-SMAL, a novel, dog-specific 3D pose and shape model derived from SMAL.
2. We develop BITE, a neural model to refine 3D dog poses by encouraging plausible ground contact, while at the same time estimating the local ground plane.
3. We show that with that model it becomes possible to reconstruct dog poses far from those encoded in a (necessarily limited) prior.
4. We advance the state of the art for monocular 3D pose estimation on the challenging *StanfordExtra* dataset.
5. We put forward a new, semi-synthetic 3D test dataset based on scans of real dogs, with which we hope to encourage a move to true 3D evaluation.

2. Related Work

There is a vast literature on 3D reconstruction of humans, while little work exists on animals. We review here model-based methods that estimate 3D animal shape and pose,

[1] In fact, recent work on 3D humans deals with advanced aspects like clothing [36, 37] or multi-person scenarios [10].

model-free methods that recover 3D shape, and relevant literature on the use of contact information, so far exploited only for humans.

2.1. Model-based 3D animal reconstruction

The majority of the methods for animal 3D pose and shape estimation, or 3D reconstruction, are based on the SMAL model, which is so far the only 3D parametric articulated shape model for generic quadrupeds. SMAL has been used to estimate 3D shape and pose of zebras [43] and dogs from monocular images [5, 21, 26] and RGBD data [19], while Biggs et al. apply SMAL to video of animals of different species [6]. Another articulated 3D shape model has been defined for birds [34]. The advantage of using articulated shape models is that 3D shape and pose priors can be defined for the species of interest, supporting reconstruction in ambiguous cases. Additionally, the recovered parametric shape and poses can be used for further analyses of body size, posture, motion and behavior.

2.2. Model-free methods for animals

Recently, due to the progress in model-free implicit and neural representations, several methods have been proposed to reconstruct 3D animals, mostly from video. The quality of the reconstruction, in terms of 3D shape and articulated motion, is still far from model-based methods. With CMR, Kanazawa et al. [18] learn to reconstruct 3D birds in a mesh-based representation, without assuming an existing bird template. LASR [39] reconstructs arbitrary 3D articulated deformable objects from video, exploiting analysis-by-synthesis. The method is highly flexible, but results are not always of good quality. BANMo [40] reconstructs animated 3D shapes from video while also learning the skeleton structure. These skeletons vary between animals, making it hard to compare motions of different individuals. TAVA [22] learns a 3D articulated model from multiple videos and a skeleton of an animal; it also learns skinning weights. Results are reported only on noise-free synthetic images, with clean backgrounds, obtained from rigged graphics models. LASSIE [41] is an optimization method that reconstructs 3D articulated animals given a set of images and a skeleton. It can be applied to generic animals, but cannot deal with occlusions and complex poses.

2.3. Contact-based methods for humans

Several recent methods use contact between a subject and the scene to improve 3D human pose estimation [15, 25, 28, 31, 35, 42] and 3D hand pose estimation [7, 11, 14, 16]. Often, methods only consider the feet [25, 42], frequently using mocap data to estimate foot contact on the ground. PROX [15] exploits contact with objects, but assumes the 3D scene is known. BEHAVE [4] uses RGB-D camera to capture humans interacting with objects and scenes.

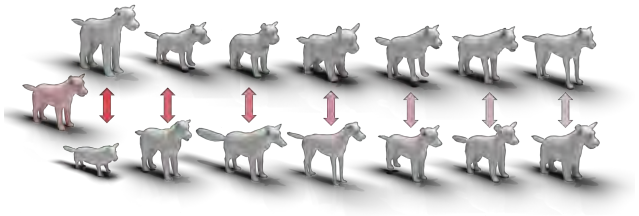

Figure 2. *D-SMAL shape space.* Shown are the mean shape and the 7 principal modes of deformation.

Very few methods use image evidence to predict contact points. HOT [8] regresses 2D contact heatmaps and body-part labels from RGB images. IPMAN [31] exploits interpenetration of the body-mesh with the ground-plane to estimate body-ground pressure from images. Fieraru et al. [12] and TUCH [24] learn to predict human self contact. RICH [17] is a dataset with contact points defined on a 3D human body. This annotation is obtained with an expensive and complex capture, requiring multi-view video and 3D scanning of subjects and the scene. As a result, the dataset has limited variability. It is used to train BSTRO, which estimates 3D contacts on a human body from a single image. In contrast, we also define contact labels on the 3D body but through an annotation procedure applied on arbitrary *in-the-wild* images, providing a richer dataset. We only exploit contact between the subject and a flat ground plane. Unlike humans, dogs do not have hands to manipulate objects, so most of their body-scene contact is with the ground.

3. Approach

3.1. D-SMAL model

Our model-based monocular reconstruction method is based on a novel, dedicated *dog-only* version of SMAL [44]. The dominant 3D parametric shape model for quadrupeds, SMAL, covers inter- and intra-species variations across a number of mammal species. It was learned from a set of toy figurines and unifies all training species (ranging from domestic cats to hippos) in a linear single shape space. When modeling on a particular species, the generic representation has undesirable properties: on the one hand the shape space is not constrained enough to prevent undesired, unrealistic shapes (e.g., a dog should not ever have the proportions of a hippo or a horse), on the other hand the shape space is too limited and not expressive enough to represent subtle, but important species-specific variations (e.g., the shapes and articulations of different dog's ears).

To train D-SMAL we employ scans of 39 toy figurines of animals in the canine family (5 used in the original SMAL model and 34 new ones). We follow the GLoSS registration procedure [44] to register a template mesh to all scans. For dog breeds with floppy ears, we set the 3D rotation of the

ears in the GLoSS template accordingly. Moreover, we also indicate whether the dog's mouth is closed (in which case we disregard mesh-to-scan distances inside the mouth when computing the loss), and define an additional lip-matching loss. Once the scans have been registered, we follow [44] to learn the shape space of D-SMAL, which we then scale to unit size and augment with variable limb lengths to increase expressiveness, as previously done when modeling dogs [5, 26]. Additionally, given that we found a few toy figurines with back and front legs of different length, we rescale the limbs such that they are more naturally proportioned. Figure 2 displays the first 7 components of our model's shape space.

3.2. BITE Network

Our network extends BARC; i.e., it takes the pose estimate from BARC, as well as the associated 2D predictions of keypoint locations and the silhouette, as intermediate results that are refined by BITE's refinement stage. The entire mapping from a raw input image to the refined 3D pose estimate forms an integrated neural pipeline.

Architecture: Figure 3 depicts the complete BITE architecture. In the first step, an input image is fed to a network derived from BARC [26]. BARC+ extends BARC by using D-SMAL and the network consists of a stacked hourglass that predicts 2D keypoints and a segmentation mask, followed by shape and pose prediction branches, where the former predicts 3D shape parameters and the camera viewpoint (translation and focal length) and the latter predicts pose parameters. BITE adds a refinement stage to increase the accuracy of the BARC+ predictions. To that end, the predicted 3D mesh is projected to the image, giving rise to 2D keypoints and a silhouette. These are combined with the intermediate keypoint and silhouette predictions from BARC+ and with the image itself, thus implicitly specifying the residual error of those intermediate predictions. All these intermediate results are fed into a 2D encoder, followed by fully connected heads that output refined 3D pose and camera parameters. We have also tried to refine the shape parameters, but empirically that proved to be neither necessary nor beneficial. For images in which the dog is on flat ground, refinement is supervised with a loss that encourages an anatomically and physically consistent placement of the 3D model on a local ground plane. In addition, further heads of the refinement stage predict per-vertex ground contact labels and a binary label that indicates whether the ground is flat. This information is collected to drive BITE's optional test time optimization (see below).

Predicting novel poses far from the prior: A subtle, but practically important bottleneck when modelling 3D animals is the limited expressiveness of the pose prior. Since the training data is insufficient to fully characterize the space of allowable poses, it is common practice to encourage predictions that are close to a small set of training poses – in the case of BARC this is through a normalizing flow prior that is trained on a set of simple poses (standing, walking) from the RGB-D dataset [19]. The prior ensures that the predictions are in a plausible region of the pose space, but at the same time rules out many realistic poses that are not represented adequately by the training data. For dogs, this concerns, in particular, sitting and lying postures, since there is, to our knowledge, no training database that adequately covers them. BITE can be understood as a means to expand the space of allowable poses, by injecting another type of evidence, namely the physical contact with the ground. In this view, our network makes the most out of BARC's restrictive pose prior, but then goes beyond it: first, it produces an imperfect, but viable estimate within the scope of the pose prior. Then, it upgrades that initial estimate to a complex pose outside the prior's domain, guided by the ground contact information (in conjunction with local joint angle constraints).

Ground contact: The refinement network predicts not only refined pose parameters, but also has an auxiliary head to estimate ground contact labels on a per-vertex basis. This head, based on a variant of the GraphCMR network with an encoder-decoder structure with skip connections [9, 20, 21], provides the detailed contact information needed for BITE's test-time optimization; see Section 3.4. Even though the assumption of locally flat ground is valid for most images, there can be exceptions. These are handled by another head of the refinement network that predicts a binary classification label for the flatness of the surface under the dog.

3.3. BITE Network Losses

BARC+ Losses: For the initial fitting stage we use the original losses proposed for BARC: reprojection losses on keypoints and the silhouette, priors for pose, camera focal length and translation, and breed losses (with the 3D breed loss adjusted to the new D-SMAL model).

Ground Contact Loss: As we do not know the location of the ground plane in the camera coordinate system, we learn to find it jointly with the dog's ground contact, or more precisely: we infer it from the dog's pose and contact points.

The topology of our dog model (which it inherits from SMAL) is such that vertices are denser on the paws and the face than on other body parts. Thus, we apply the ground contact loss to a uniformly meshed surface, obtained by Voronoi clustering [3, 32, 33] on the mesh with the mean shape in the canonical "T" pose. We map the vertex-wise ground contact labels from the D-SMAL model to the new surface and aim to find a plane that minimizes the squared sum of distances between the set of m contact points $(p_1, \ldots, p_m)$ and the plane. Given a point c belonging to the plane and a unit normal vector n, which determines the plane orientation, the orthogonal distance be-

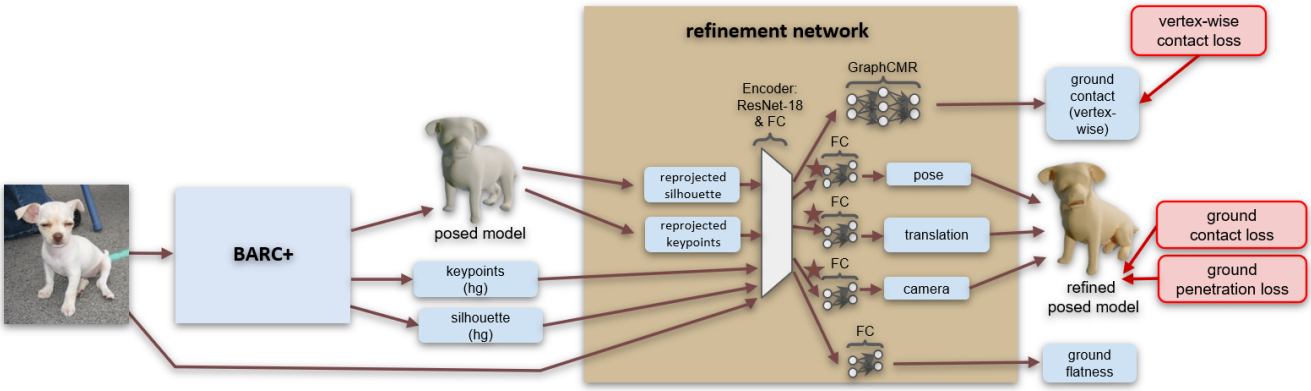

Figure 3. *BITE network.* Our network exploits ground contact constraints and refines BARC+ predictions. The brown stars denote that additional information is fed to the respective head. BARC+ refers to the network of [26] with the new D-SMAL model.

tween a point p_i and the plane is $(p_i - c)^\top n$. The plane itself can be found by minimizing:

$$\min_{c,\|n\|=1} \sum_{i=1}^{n} \left((p_i - c)^\top n\right)^2. \quad (1)$$

Solving this equation for c gives $c = \frac{1}{m}\sum_{i=1}^{m} p_i$, which corresponds to the center of the contact points. Introducing the $3 \times m$ matrix $A = [p_1 - c, ..., p_m - c]$, Eq. 1 can be rewritten as $\min_{\|n\|=1} \left\|A^\top n\right\|_2^2$ and solved by a singular value decomposition of $A = USV^\top$:

$$\|A^\top n\|_2^2 = \|S^\top U^\top n\|_2^2 = (\sigma_1 y_1)^2 + (\sigma_2 y_2)^2 + (\sigma_3 y_3)^2, \quad (2)$$

where $\sigma_1, \sigma_2, \sigma_3$ are the singular values, listed in decreasing order, and y is the unit vector $y = U^\top n$. Eq. 2 achieves its minimum for $y = (0, 0, 1)^\top$ and equivalently $n = U(:, 3)$. Plugging this into Eq. 1 leads to a minimum value of σ_3^2, which is thus equivalent to the sum of squared of distances between the contact points and the best fitting plane [2, 13, 29]. We define our ground plane loss as:

$$L_{\text{gc,plane}} = \sigma_3^2. \quad (3)$$

Ground Penetration Loss: To strengthen the ground *contact* loss, we complement it with a ground *penetration* loss that penalizes all non-contact vertices located below the ground plane. To that aim, we first evaluate if the plane normal calculated in the last step points up or down, and define a ground penetration loss for the set of points $(\widehat{p}_1, \ldots, \widehat{p}_{\widehat{m}})$ below the plane as:

$$L_{\text{gc,pen}} = \sum_{i=1}^{n} (\widehat{p}_i - c)^\top n. \quad (4)$$

Pose Regularizers for the Refinement Network: To serve its purpose, the refinement network must be able to predict poses that considerably differ from those preferred by the pre-trained pose prior. Still, it must respect anatomical limits and should only deviate from the initial, tightly constrained fit as far as necessary. We penalize sideways movements of the legs as well as their torsion. The structure of the refinement network, which predicts differential updates to the pose, makes it possible to explicitly regularize the deviations, by multiplying them by a constant factor, that we set at 0.1. There are thus no hard constraints on how far the refinement can deviate from the initial estimate, but training starts with small changes, and converges as desired. Predicting an additive correction and scaling it down is an important engineering detail, which helps the network to benefit from the overly tight pose prior without being bound by it.

Reprojection Losses: We exploit the same reprojection losses as [26], namely a reprojection loss for the keypoints and one for the silhouette.

Vertex-Wise Ground Contact: We treat the ground contact prediction as a binary classification and apply a per-vertex cross-entropy loss.

Ground Flatness: We also apply a cross-entropy loss on the label for flatness of the ground.

3.4. BITE Optimization

While a single run of BITE's refinement stage already gives very accurate predictions (see experiments), they can be improved even further with an iterative optimization loop. This is a test-time procedure driven by the individual test image, but thanks to our novel ground contact losses, we are now able to refine results based on additional image evidence without over-fitting to the image. Naturally, this optional extension comes at the cost of increased runtime.

Optimization is performed over translation, camera focal length, dog pose and dog shape. While the BITE network optimizes shape in terms of shape parameters β (consisting of PCA coefficients β_{pca} and limb length parameters κ), we do the same for the first half of the iteration steps within the test time optimization but, in the second half, allow additional symmetric vertex shifts.

Similar to the dog model in [5, 26], D-SMAL is a function $M(\beta_{pca}, \kappa, \theta)$ of PCA shape coefficients β_{pca}, limb length coefficients κ and pose θ. The steps from model parameters to the final posed mesh can be summarized as adding shape blend shapes through a linear blend shape function $B_S(\beta_{pca})$ to a template mesh T, resulting in $T_S = T + B_S(\beta_{pca})$ and in a second step deforming the mesh further by posing the model according to pose parameters. Limb length changes and translation are also applied within this second step and we describe this second step as a function $F(T_S, \kappa, \theta)$. To go beyond the shape space of the current model, we introduce a set of vertex shifts υ that are added to T_S. We end up with $M(\beta_{pca}, \kappa, \theta) = F(T_S + \upsilon, \kappa, \theta) = F(T + B_S(\beta_{pca}) + \upsilon, \kappa, \theta)$.

Dog shape estimation from an image is a highly underconstrained problem; e.g. at least half the dog is always occluded. We compensate for that weak visual cues by assuming symmetric dog shape. The vertex shifts υ are implemented as a function of a vector with three entries for each left vertex and two entries for each center line vertex. Shifts for right vertices follow directly from the shifts for the left side. Note that symmetric vertex shifts help limit over-fitting to 2D evidence but cannot model asymmetric shape details and pose dependent deformations. But such deviations from the D-SMAL model are small relative to the limited expressiveness of the shape space in modeling widely varied dog shapes from images.

Losses: The optimization loop minimizes the same reprojection losses used during network training. Since there are no silhouette and keypoint annotations available for test images, we exploit the predictions of the BARC+ stacked hourglass. The ground contact and ground penetration losses are also similar to the losses applied within the network, supervised by the predicted vertex-wise contact labels and only active if the ground plane is predicted to be flat. Furthermore, we penalize torsion as well as sideways movements of legs and tail. Those losses are augmented with three 3D regularization terms: (1) A normal consistency loss $L_{\text{3D, norm}}$ that is computed as the consistency of the normals n_1, n_2 for each pair of neighboring faces: $L_{\text{3D, norm}}(n_1, n_2) = \sum 1 - cos(n0, n1)$, (2) An edge loss $L_{\text{3D, edge}}$ that encourages a mesh with uniform edge lengths, and (3) a loss that encourages the Laplacian of the deformed mesh to be similar to the Laplacian of the mesh before vertex-shifts were allowed, see [21].

4. Experiments

4.1. Training Data

We train our network on the StanfordExtra image dataset [5]. This dataset is annotated with 2D keypoints, silhouettes and breed labels. Similarly to [26], we use the training set extended with pseudo-ground truth withers, throat and eyes keypoints. We furthermore collect vertex-wise ground contact labels by first grouping the images based on poses, visibility, and ground flatness. We then label a selection of the StanfordExtra training images with contacts either on a per-paw or per-vertex level, depending on the type of pose. The final training set has contact labels for 2554 images of standing or walking dogs, 795 lying dogs, 867 sitting dogs and 89 dogs in other poses. Those groups based on pose are re-used within our data sampler in order to have more balanced batches. By default, we train versions of our network that include ground contact and ground penetration losses with a data sampler that, for each batch, randomly selects 2 images with non-flat ground, and 12 images with flat ground. Out of those 12 images, 2 show sitting poses, 2 lying poses, 2 poses of dogs moving or standing with less than 4 paws on the ground, 4 images of dogs standing on all paws and 2 images that belong into neither of those groups. Within an ablation study, we show that this data sampling strategy helps, but it is not crucial and can be avoided in case labels of this kind are not available.

4.2. Evaluation Protocols

Due to the lack of ground truth 3D data, previous work has typically relied on 2D reprojection errors. For such an evaluation, the predicted 3D model is projected into the image and an intersection over union (IoU) score for the silhouette as well as a score measuring the percentage of correct keypoints (PCK) are reported. For good predictions, those scores should to be high, but high reprojection scores do not guarantee an accurate 3D estimate due to ambiguities in depth. Current methods may be overfit to these metrics. Still, for completeness, we report IOU and PCK scores in the supplementary material.

To enable quantitative evaluation of 3D shape, we contribute a new test dataset with 3D ground truth shapes. We purchase 26 textured scans of six dogs: a Labrador, a Bulldog, a Husky, an Akita, a Malinois and a Dalmatian. We scale the dog scans such that they have realistic relative size, place each of them on a virtual lawn to model basic occlusions that often happen around the paws, and render 7 images per dog from different viewpoints. The rendered images have size of 1080×1080 pixels, but we use ground truth bounding boxes to prepare the input for all networks to test. Examples of the rendered images can be found in Fig. 7.

We manually label a small set of easy-to-locate keypoints on the scans and on the D-SMAL model and use those to initialize a rigid alignment (rotation, translation, and scaling) between prediction and ground truth scan. The rigid alignment is refined based on the absolute distance between each scan vertex and the closest point on the surface of the predicted mesh (scan-to-mesh distance). This evaluation follows the evaluation scheme proposed as part of the

NoW benchmark for faces [27].

In contrast to faces, animal bodies show complex deformations and only measuring the distances in one direction, namely from scan to the predicted mesh, can understate errors, as the scan only represents the outer surface of the dog; occluded surfaces are not represented. Therefore, we complement the evaluation with mesh-to-scan distance. To calculate this, we apply the same uniformly re-meshed D-SMAL surface used to calculate the ground contact loss.

4.3. Comparison to Related Work

We compare BITE and BITE-ttopt (*i.e.*, including test time optimization) to WLDO [5], coarse-to-fine animal pose and shape estimation (CTF) [21], and BARC [26]. Errors are reported in Table 1. The evaluation shows that we outperform all previous methods. We show qualitative examples in Fig. 4 and Fig. 5; further illustrations are in the supplementary material.

	error [cm]	
	scan→mesh	**mesh→scan**
WLDO [5]	2.65	7.55
CTF [21]	2.59	6.17
BARC [26]	2.40	3.93
BITE (ours)	2.07	3.15
BITE-ttopt (ours)	**2.03**	**2.84**

Table 1. Comparison to SOTA.

	error [cm]	
	scan→mesh	**mesh→scan**
BARC+ w/o gc	2.32	3.92
BARC+ with gc	2.13	3.48
BITE w/o gc	2.30	4.16
BITE w/o sampler	2.09	3.31
BITE with shape in BARC+	2.12	3.17
BITE with shape in ref	2.29	4.24
BITE (ours)	2.07	3.15
BITE-ttopt (ours)	**2.03**	**2.84**

Table 2. Ablation study.

4.4. Ablation Study

First, we train the BARC network, with the new D-SMAL model. We call this network BARC+. We do not change the network or the losses, but only adjust the breed losses to the new model and make a few minor changes, such as re-adjusting the dog's toe keypoint. This experiment is called *BARC+ w/o gc*, where "gc" stands for ground contact. We show that the new dog model leads to a small improvement over BARC in terms of scan-to-mesh and mesh-to-scan errors, see Tab. 2 *BARC+ w/o gc* compared with *BARC* in Tab. 1, but also to a visible improvement in terms of perceptual results, see supplementary material.

Next, we examine the full network training. Table 2 shows results for our method excluding test time optimization (*BITE*) and including test time optimization (*BITE-ttopt*). If we forgo the data sampling procedure described in Sec. 4.1, we obtain slightly worse results. When not exploiting our novel ground contact losses (*BITE w/o gc*) within the refinement network, the results are significantly worse. The refinement network in that case overfits to the image evidence, and even though IOU and PCK scores on the StanfordExtra test set are good, our 3D evaluation reveals significant deviations from the true pose and shape. The same holds when looking at the BARC+ part of the network and comparing the experiment with ground contact losses (*BARC+ with gc*) to the one without (*BARC+ w/o gc*).

Another important observation is that, if we use ground contact losses, the refinement network leads to a significant boost in performance (*BITE vs. BARC+ with gc*). The challenge is to exploit an existing pose prior (in our case based on Normalizing Flows), but still be able to significantly deviate from what it has seen at training time. Some poses, like standing and symmetric sitting poses, such as the left-most example in Figure 6, are close enough to the prior's expectation to be reconstructed with the BARC+ network augmented with ground contact losses. But totally unseen poses, like the right two images in Figure 6, require a network with more flexibility.

We also conduct experiments where we optimize the shape during the refinement step as well. We have tested two settings, either adding a head for shape estimation to the refinement branch or allowing gradients to back-propagate to the shape branch of the BARC+ part. Both variants degrade 3D performance, due to undesired shape changes when trying to fulfil the ground contact constraints. To see this, think how shrinking the legs instead of lying down can bring the torso to the ground. Outsourcing shape prediction to the training of the BARC+ part stabilises the shape predictions, since the breed losses help to prevent bad shape predictions even when faced with inaccurate poses. We do, however, allow shape changes during test time optimization to obtain the best result.

5. Conclusion

We present a method for 3D dog reconstruction from monocular images. The 3D pose and shape estimation of articulated bodies from single images is a challenging problem that, for dogs, has been so far poorly addressed, due to issues like self-occlusion, no notion of depth, restrictive pose priors and insufficiently expressive shape spaces. Even though attempts have been made to go beyond the shape spaces of available dog models by predicting vertex shifts, those methods are not yet capable of predicting realistic 3D dog meshes. We move forward by providing a new

Figure 4. *Comparison with SOTA methods.* This figure shows WLDO [5], CTF [21], BARC [26] and our results.

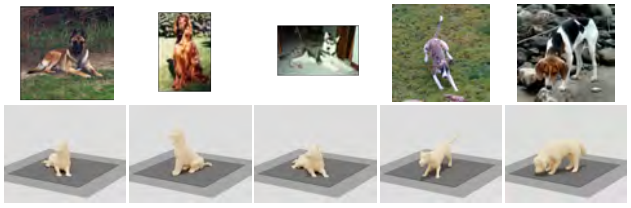

Figure 5. *BITE results.* Visualization in 3D illustrates the ground contact.

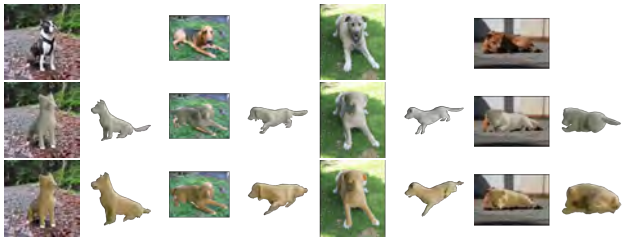

Figure 6. *Benefit of the refinement network.* Top row: input image, Middle: BARC+ with gc, bottom row: BITE. The figure illustrates the importance of the refinement stage: the pose prior is limited in its ability to predict poses outside the training distribution.

dog model called D-SMAL, with an extended shape space. Then, we tackle pose estimation. In contrast to humans, pose priors for animals are very limited. Instead of capturing 3D poses of dogs, which is hard, we take an alternative approach, and attempt to learn to generate complex, self-occluded 3D poses by enriching our knowledge about the 3D image content. Specifically, we introduce ground contact constraints and design a network such that we can still

Figure 7. Renderings of scanned dogs serve as test images.

exploit old pose priors to initialize the predictions, but afterwards deviate from those values and, if necessary, go far beyond the original prior's target region. To the best of our knowledge, we, for the first time, predict realistic complex poses such as sitting and lying postures for dogs. After estimating reasonable shape and pose parameters, we are able to go beyond the D-SMAL shape space by allowing free vertex shifts. Previous work on 3D dog pose has been evaluated in 2D, but IOU and PCK scores are not good indicators of 3D quality. We contribute with a novel dataset with 3D ground-truth, obtained by rendering real dogs captured in 3D. Our BITE method outperforms all previous work on this dataset. Future work should consider self-intersections, use inferred shapes to expand the D-SMAL shape space, and learn more complex pose-dependent deformations.

Acknowledgments. 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_2023.txt

References

[1] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *SIGGRAPH*. 2005. 1

[2] K Somani Arun, Thomas S Huang, and Steven D Blostein. Least-squares fitting of two 3-D point sets. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1987. 5

[3] M Audette, D Rivière, M Ewend, A Enquobahrie, and S Valette. Approach-guided controlled resolution brain meshing for fe-based interactive neurosurgery simulation. In *Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshops*, 2011. 4

[4] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2022. 3

[5] 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 *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 4, 6, 7, 8

[6] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and SMAL: Recovering the shape and motion of animals from video. In *Asian Conference on Computer Vision (ACCV)*, 2018. 3

[7] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[8] Yixin Chen, Sai Kumar Dwivedi, Michael J. Black, and Dimitrios Tzionas. Detecting human-object contact in images. June 2023. 3

[9] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *European Conference on Computer Vision (ECCV)*, 2020. 4

[10] Zijian Dong, Jie Song, Xu Chen, Chen Guo, and Otmar Hilliges. Shape-aware multi-person pose estimation from multi-view images. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[11] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. June 2023. 3

[12] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3D human self-contact. In *Conference on Artificial Intelligence (AAAI)*, 2021. 3

[13] Walter Gander and Jiri Hrebicek. *Solving problems in scientific computing using Maple and Matlab®*. Springer Science & Business Media, 2004. 5

[14] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2021. 3

[15] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[16] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019. 3

[17] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2022. 3

[18] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[19] Sinéad Kearney, Wenbin Li, Martin Parsons, Kwang In Kim, and Darren Cosker. RGBD-Dog: Predicting canine pose from RGBD sensors. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2020. 3, 4

[20] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019. 4

[21] Chen Li and Gim Hee Lee. Coarse-to-fine animal pose and shape estimation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3, 4, 6, 7, 8

[22] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. TAVA: Template-free animatable volumetric actors. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 1, 2

[24] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2021. 3

[25] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[26] Nadine Rueegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. BARC: Learning to regress 3D dog shape from images by exploiting breed information. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2022. 2, 3, 4, 5, 6, 7, 8

[27] 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 *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019. 7

[28] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. PhysCap: Physically plausible monocular 3D motion capture in real time. *ACM Transactions on Graphics (TOG)*, 39(6), 2020. 3

[29] Inge Soederkvist. Using SVD for some fitting problems. https://www.ltu.se/cms_fs/1.51590!/svd-fitting.pdf. 5

[30] Jacob D. B. Stillman and Eadweard Muybridge. *The horse in motion as shown by instantaneous photography, with a study on animal mechanics founded on anatomy and the revelations of the camera, in which is demonstrated the theory of quadrupedal locomotion*. J. R. Osgood and Company, Boston, 1882. 1

[31] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[32] Sébastien Valette and Jean-Marc Chassery. Approximated centroidal Voronoi diagrams for uniform polygonal mesh coarsening. In *Computer Graphics Forum*, 2004. 4

[33] Sébastien Valette, Jean Marc Chassery, and Rémy Prost. Generic remeshing of 3D triangular meshes with metric-dependent discrete Voronoi diagrams. *Transactions on Visualization and Computer Graphics (TVCG)*, 2008. 4

[34] Yufu Wang, Nikos Kolotouros, Kostas Daniilidis, and Marc Badger. Birds of a feather: Capturing avian shape models from images. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2021. 3

[35] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. CHORE: Contact, human and object reconstruction from a single rgb image. *arXiv preprint arXiv:2204.02445*, 2022. 3

[36] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal Integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[37] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. ICON: Implicit clothed humans obtained from normals. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2022. 2

[38] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & Ghuml: Generative 3D human shape and articulated pose models. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2020. 1, 2

[39] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. LASR: Learning articulated shape reconstruction from a monocular video. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2021. 3

[40] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. BANMo: Building animatable 3D neural models from many casual videos. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2022. 3

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

[42] Yuliang Zou, Jimei Yang, Duygu Ceylan, Jianming Zhang, Federico Perazzi, and Jia-Bin Huang. Reducing footskate in human motion reconstruction with ground contact constraints. In *Winter Conference on Applications of Computer Vision (WACV)*, 2020. 3

[43] 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 *International Conference on Computer Vision (ICCV)*, 2019. 3

[44] Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2017. 1, 2, 3, 4