RGBT-Dog: A Parametric Model and Pose Prior For Canine Body Analysis Data Creation

Jake Deane[1] Sinéad Kearney[1] Kwang In Kim[2] Darren Cosker[1]
University of Bath[1] POSTECH[2]
jcdeane21@gmail.com, sineadkearney08@gmail.com, kimkin@postech.ac.kr, dpcc22@bath.ac.uk

Abstract

While there exists a great deal of labeled in-the-wild human data, the same is not true for animals. Manually creating new labels for the full range of animal species would take years of effort from the community. We are also now seeing the emerging potential for computer vision methods in areas like animal conservation, which is an additional motivation for this direction of research. Key to our approach is the ability to easily generate as many labeled training images as we desire across a range of different modalities. To achieve this, we present a new large scale canine motion capture dataset and parametric canine body and texture model. These are used to produce the first large scale, multi-domain, multi-task dataset for canine body analysis comprising of detailed synthetic labels on both real images and fully synthetic images in a range of realistic poses. We also introduce the first pose prior for animals in the form of a variational pose prior for canines which is used to fit the parametric model to images of canines. We demonstrate the effectiveness of our labels for training computer vision models on tasks such as parts-based segmentation and pose estimation and show such models can generalise to other animal species without additional training.

1. Introduction

Humans are the focus of many challenges in computer vision [9, 14, 15, 37, 38, 47] but what about animals? Like their human counterparts, training deep learning based approaches for animal computer vision requires a great deal of data. However, animals display significant inter and intra-species variation in appearance. This makes the task of labelling data more difficult than it is for human subjects - especially if we wish to cover large ranges of animal species. While some animal data is available [5, 8, 11, 19, 49] it is generally sparse both in terms of quantity and modality or is missing important semantic information. A number of these sets are also labelled by domain adaptation methods and may contain inaccurate labels. We may also wish to have data such as part-segmentation labels or 3D keypoints that are not provided by existing datasets. We therefore ask: Is it feasible to avoid manual labeling in these situations?

We demonstrate (*Contribution 1*) that it is possible to perform difficult computer vision tasks for canines – focusing on parts-based segmentation and dense joint detection to illustrate – using only synthetically generated data labels. We focus upon canines due to the large degree of inter-species variation presented by dogs. To achieve this, (*Contribution 2*) we propose an approach to generate multi-task data labels for canine images in a wide range of data poses leveraging a new canine parametric body model, RGBT-Dog, with a large compatible motion capture library of canine motion which allows for realistic posing and dynamic motion generation. This model also comes with the first variational pose prior for canines (*Contribution 3*) and, to the best of our knowledge, animals in general. This approach allow us to generate synthetic labels for (1) real in-the-wild images as well as (2) fully synthetic images and labels. We also demonstrate (*Contribution 4*) that models trained using these data-points can generalise to other animals. We release our parametric model, networks, motion capture, and image labels as part of our work.

2. Related work

Visual analysis of animals. Animals are part of large image classification and segmentation datasets such as COCO [26] and OpenImages [22, 23] but there has been little work into animal body analysis compared to their human counterparts due to an absence of data. The TigDog [11, 12], StanfordExtra [5] and BADJA [6] datasets for 2D pose estimation were recently introduced. Yu *et al*. [45] introduced a large dataset of multi animal pose data though the number of keypoints is somewhat limited due to lacking certain keypoints such as those in the tail. Mathis *et al*. introduced DeepLabCut [24, 29, 30] for automatically labelling 2D keypoints which was extended by Nath *et al*. [32] for 3D pose.

Shooter *et al*. [41] introduced a synthetic dataset of dogs for 2D pose estimation.

Biggs *et al*. [6] predict 2D keypoints for recovering 3D shape and motion of animals while Zuffi *et al*. [49] introduced a method for estimating 3D pose, shape and texture from in the wild images of Grevy's zebras. Similarly Biggs *et al*. [5] introduced a method for recovering the 3D shape and pose of dogs, utilising a rich prior over shapes producing 2D keypoints for the Stanford dogs dataset. In contrast Cao *et al*. [8] take advantage of the large amount of human data available via cross-domain adaptation. Li and Lee [25] also use domain adaptation, focusing upon closing the gap between real and synthetic data for 2D animal pose estimation. For body pose estimation, [8, 25, 31] present joint keypoints for different animal classes but they are limited in number and ignore keypoints relevant to animals such as those in the tail. Biggs *et al*. [5] provide additional keypoints but their data is limited to key-points and silhouette maps. To our knowledge, no dataset of 3D animal keypoints for animals exists nor are there dense part segmentation datasets or dense keypoint labels – something we address in our work. Also addressed is the limited compatible motion capture data with prior models, making the use of realistic poses upon which to generate data difficult.

Parametric human and animal models. Parametric models have been commonly used in human body shape and pose analysis. SMPL [27] is one of the best-established models with respect to human body modelling, improving upon the SCAPE model [3]: SMPL uses a set of shape and pose parameters to generate skinned vertex based models in a wide variety of poses, and can be used to generate synthetic images and data. A number of works have focused upon recovering these parameters from 2D RGB images such as Madadi *et al*. [28] and Kanazawa *et al*. [18] which aim to regress the shape and pose parameters of SMPL from RGB images directly. Two important works in this area are SMPLify by Bogo *et al*. [7] and SMPLify-X by Pavlakos *et al*. [36]. The former introduced a method for capturing 3D human shape and pose from an image via 2D points obtained from a convolutional neural network upon which SMPL was fit and optimised. Pavlakos *et al*. [36] expanded upon this by introducing SMPL-X and SMPLify-X, while a more recent extension is STAR [34].

There is significantly less research on the parametric modeling of animal body shapes: Animals are less cooperative than humans and more logistically difficult for data capture with respect to obtaining 3D scans. There is also a significantly greater shape variation among animals compared to people. Zuffi *et al*. [48] introduced the SMAL model, based upon 3D scans of deformable animal toys. Zuffi *et al*. [50] later introduced SMALR enabling capture of greater detailed 3D shape and the modelling of new species while also allowing for the extraction of texture maps to create textured meshes. Both Biggs *et al*. [6] and Rüegg *et al*. [39] introduce scaling parameters to the SMAL model to allow for breed specific fitting for dogs. However, both of these methods do not change the underlying model.

While existing models could also be potentially used to generate synthetic labels for computer vision tasks, they are not compatible with large scale existing motion data already available in the community [19]. Our new parametric model, RGBT-Dog, greatly builds on this data set, expanding the number of possible synthetic poses and thus labels that can be generated and used to train vision tools for (e.g., without data of animals in certain poses, you cannot train accurate pose estimators for these situations). In addition, being generalised models they are not as anatomically correct as our canine specific model and allow for less canine specific labels (such as those of the tail). Our model also provides an increased number of skeletal keypoints; 43 compared to the 33 of the SMALR model and its derivatives. Our model, unlike previous works is also compatible with a large scale motion capture dataset and an associated variational pose prior which has not been possible for previous work.

Synthetic data for task learning. Synthetic data has seen an increased use recently; In 3D human pose estimation [10, 40, 42, 43], ground truth data is difficult to produce and they also suffer from a lack of statistical variety due to images being captured in indoor environments. The SURREAL dataset of Varol *et al*. [43] used the SMPL model and motion capture data to produce synthetic RGB images and data labels. Wood *et al*. [44] demonstrate it is possible to perform in the wild face related computer vision using synthetic data. For animal tasks, Mu *et al*. [31] introduced a semi-supervised learning method trained upon real data of humans and synthetic data produced from CAD animal models. Fangbemi *et al*. [13] present a method for creating synthetic images from 2D videos with 2/3D keypoints. Our dataset improves upon existing datasets by providing dense keypoints labels for 2D and 3D where only sparse labels had been available before while also providing labels for other tasks such as dense part-segmentation.

3. RGBT-Dog

To achieve the generation of multi-modal data labels for canine images we present a new parametric model for canine shape and texture - RGBT-Dog - that can be posed realistically by utilising an extended set of new canine motion capture data. We use this model to generate labeled data in two ways (1) create new synthetic dog images and labels (any modality from the model, e.g. dense landmarks, depth maps, segmentation maps), and (2) create synthetic labels for real images from existing data sets. We achieve

(2) by bootstrapping our model on the canine keypoint data provided by [5]; these keypoints are much sparser than we desire (only a small set of key points compared to RGBT-Dog's), though they can be used to help fit RGBT-Dog to these real images and then propagate as rich a set of labels on this data as we require.

RGBT-Dog shares some similarities to SMAL and SMALR [48, 50] as a parametric model. Like other parametric models [27, 35, 36, 48, 48] RGBT-Dog is defined by a set of parameters. These parameters are used to construct an articulated 3D mesh model with $N = 2,426$ vertices and $K = 43$ joints via vertex based linear blend skinning. RGBT-Dog is paramaterised by a set of shape $\beta \in \mathbb{R}^{10}$, pose $\theta \in \mathbb{R}^{43\times3}$, texture $\gamma \in \mathbb{R}^{11}$, root translation $t \in \mathbb{R}^3$, root rotation $r \in \mathbb{R}^3$ parameters, and skinning weights w.

To give an abridged summary of RGBT-Dog; β deforms the template mesh to produce a body shape which is then manipulated into a pose via θ. This creates displacements from the template vertices where we then apply a standard linear blend skinning function $\mathcal{W}(\cdot)$ using skinning weights w. This creates an articulated mesh. A UV texture map for this mesh is generated using γ and applied to the mesh to give it a texture. r and t are used to position the textured mesh in 3D space. This creates our textured, posed 3D canine mesh $m \in \mathbb{R}^{3\times2426}$. This overall process is summarised by the model function $Q(\cdot)$ (Eq. (5)).

3.1. Shape

Like many parametric models we utilise a PCA shape space to build a mesh from the template. We build upon the shape space of Kearney *et al.* [19] where we use the same meshes for 11 real dogs which have been further refined by a digital artist for greater detail. As a result our shape parameter, β is the same dimensonality as that of [19].

3.2. Texture

Our PCA texture space is generated from the texture maps of 12 UV scans (Fig. 1). Each texture map is originally represented as a multi-dimensional array: $\{T_i\}_{i=1}^{12} \subset \mathbb{R}^{f\times d\times d\times d\times 3}$ where $f = 4,848$ is the number of mesh faces and $d = 4$ is the resolution of the texture, and we convert it to a vector of size $f \times d \times d \times d \times 3$. Each element of T_i is normalized into the range $[0, 1]$. Applying PCA to $\{T_i\}$, we obtain the eigenvector matrix $E = [\mathbf{e}_1, \dots, \mathbf{e}_{12}]^\top$ with normalized eigenvectors $\{\mathbf{e}_i\}_{i=1}^{12}$ of the covariance matrix of $\{T_i\}$. Given this model, a new texture T' can be generated using the first 11 principal components of E by

$$T' = \tau(E\gamma + \overline{T}), \tag{1}$$

where $\overline{T}$ is the mean texture, $\gamma \in \mathbb{R}^{11}$ is a randomly sampled vector, and τ is a threshold operator confining the outputs to be in the range $[0, 1]$. Some principal components are visualised on the left of Fig. 2.

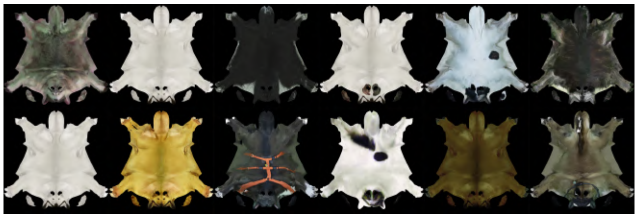

Figure 1. Dog textures as UV maps produced from cleaned photogrammetary scans.

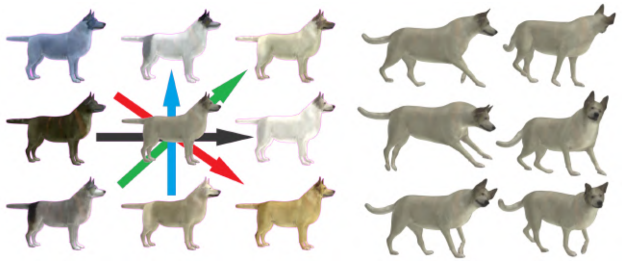

Figure 2. Left: First four principal components of the texture PCA space, displayed on a generic dog mesh. The mean texture is in the middle, with the red, green, blue and grey arrows representing the first, second, third and fourth components respectively. Each component shows ±2 standard deviations. Right: Motion from [19] applied to the dog with mean texture.

3.3. Pose, motion capture and variational pose prior

The skeleton of RGBT-Dog is described by a hierarchy of 43 joints defined by a kinematic tree which keeps the parent/child relation of each joint where the root is the base of the spine. The joint hierarchy is represented by a 43×3 matrix, θ, which corresponds to the relative rotations from parent joints where each rotation is expressed in an axis angle format of 3 values, $\rho \in \mathbb{R}^3$.

Unlike our shape and texture parameters, we cannot just randomly generate a set of pose parameters to pass to the RGBT-Dog model ($Q(\cdot)$) due to the hierarchical natures of the axis angle rotations with respect to the canine skeleton; every joint keypoint location is based upon the position of its parent keypoint (e.g. the position of the foot relies upon the position of the ankle). If we were to randomly generate pose parameters, we would obtain dog skeletons posed in impossible positions such as backward facing knee joints or the neck turned at impossible angles. Instead, like in SURREAL [43], we use motion capture data.

We introduce a new motion capture dataset of 11 dogs performing a variety of actions, which we can use to pose the RGBT-Dog mesh m. When combined with the data from Kearney *et al.* [19] (which contained 5 dogs) this results in the largest publicly available data set of dog motion capture. For capture we used a 20-camera Vicon motion capture system with processing using Vicon's Shogun soft-

Table 1. Statistics of our motion capture data with the number of frames given per motion. 'Tab' is shorthand for Table. '-' indicates that the corresponding motion was not recorded for a given dog.

	Motion						
Dog	Walk	Trot	Jump	Poles	TabOn	TabOff	TabOnOff
Dog 1	756	262	610	368	184	440	-
Dog 2	528	388	466	460	-	-	1,198
Dog 3	1,252	-	768	-	-	-	-
Dog 4	1,279	456	395	541	-	-	1,068
Dog 5	408	-	218	388	-	-	804
Dog 6	328	372	260	440	-	-	740
Dog 7	548	436	336	488	-	-	338
Dog 8	512	428	240	280	351	307	-
Dog 9	996	292	292	1,068	-	-	744
Dog 10	904	228	460	900	-	308	-
Dog 11	596	234	440	744	300	352	-
Dog 12	640	576	260	1,280	440	332	-
Dog 13	440	189	227	260	527	244	-
Dog 14	299	271	240	460	-	-	620
Dog 15	496	372	287	452	356	188	-
Dog 16	685	228	248	539	-	-	460

ware and a in-house solving skeleton. The 11 new dogs performed the same motions and procedure as outlined in Kearney *et al*. [19] to ensure parity and these motions can be seen in Tab. 1.

We use this motion capture data to create the first, to our knowledge, variational canine pose prior. Our variational canine pose prior is inspired by that of SMPLify-X [36]. As in human pose we wanted a prior that allowed feasible poses while penalising impossible ones (e.g. legs turned round the wrong way). As the prior of [36] was able to accomplish this goal for humans we elected to replicate it for canines. As in [36] we trained our pose prior using a VAE to learn a latent representation of canine poses. Compared to [36] who used pose parameters from large human motion capture datasets [1, 2, 16], we use a comparatively smaller set of canine motion data which is detailed above. We extract pose parameters from this motion capture data via inverse kinematics. The motion data for dogs 3 and 10 (see Tab. 1) are used as the test data. All other dogs are used for training across all motions. Our VAE architecture follows that of [36] with respect to the number and type of layers employed but note that we do change the size of some hidden layers to accommodate our larger number of skeleton joints (43 to SMAL's 23). We also use a latent dimension of 40 as opposed to the 32 in [36].

Where we differ is in the choice of our loss function, which consists of four distinct terms: A Kullback-Leibler loss ($\mathcal{L}_{KL}$), a mesh reconstruction loss ($\mathcal{L}_{mesh}$), a pose reconstruction loss ($\mathcal{L}_{pose}$) and a regularisation loss ($\mathcal{L}_{reg}$). $\mathcal{L}_{mesh}$ is the loss between the vertices of a mesh posed with the pose parameters (θ), $v \in \mathbb{R}^{2426 \times 3}$ and the vertices of the mesh posed by the decoder's output, $\hat{v} \in \mathbb{R}^{2426 \times 3}$. Likewise $\mathcal{L}_{pose}$ is the reconstruction loss between the input pose in the form of the rotation matrix, $R \in SO(3)$ and decoder output rotation matrix, $\widehat{R}$. $\mathcal{L}_{reg}$ is a regularisation loss on the VAE network weights, κ, to avoid large weights. $\mathcal{L}_{KL}$, as in Pavlakos *et al*. [36] is a Kullback-Leibler term that follows the formulation in [21] where $Z \in \mathbb{R}^{40}$ is the latent space of the VAE model. These losses are combined in $\mathcal{L}_{total}$ (Eq. (2)) where $c_1 = 0.005$, $c_3 = 0.0001$ and $c_2 = 1 - c_1$ if the number of epochs is less than ten and zero otherwise. The last terms were set so that the learned VAE generates valid poses but does not over-train to reproduce the same pose.

$$\mathcal{L}_{total} = c_1 \mathcal{L}_{reg} + c_2 \mathcal{L}_{mesh} + c_2 \mathcal{L}_{pose} + c_3 \mathcal{L}_{reg}, \tag{2}$$

$$\mathcal{L}_{KL} = KL(q(Z|R)||\mathcal{N}(0, I)),\ \mathcal{L}_{reg} = \|\kappa\|^2, \tag{3}$$

$$\mathcal{L}_{pose} = \|R - \widehat{R}\|,\ \mathcal{L}_{mesh} = \|v - \hat{v}\|_2^2, \tag{4}$$

where KL and $\mathcal{N}$ represent the Kullback–Leibler divergence and Gaussian distribution, respectively.

4. Data creation

Our RGBT-Dog model, Q, uses standard vertex-based linear blend skinning to generate a mesh m with $N = 2,426$ vertices and $K = 43$ joints:

$$m = Q(\beta, \gamma, \theta, r, t, w), \tag{5}$$

where shape $\beta \in \mathbb{R}^{10}$ and texture $\gamma \in \mathbb{R}^{11}$ parameters are used to explore the respective mesh shape and texture PCA spaces. $\theta \in \mathbb{R}^{43 \times 3}$ are the pose parameters presented in Euler axis angle of the model skeleton joints. Root rotation $r \in \mathbb{R}^3$ and translation $t \in \mathbb{R}^3$ are used to determine the dogs position and orientation in 3D space. Blend weights $w \in \mathbb{R}^{N \times K}$ are used with a standard linear blend skinning function to skin the mesh via the deformed vertices. Finally, the texture map generated from our texture parameter γ is applied to the canine mesh creating the final textured, posed canine mesh m. A detailed overview of this process can be found in the supplementary material. Once constructed, a mesh m is rendered using a renderer Π:

$$[I, K_{2D}, K_{3D}, P, S] = \Pi(m, \Gamma) \tag{6}$$

with a set of rendering parameters Γ. Our camera possesses no intrinsic rotation and translation as these are directly passed to the root of m via r and t respectively. This produces an image I, 43 3D keypoints K_{3D}, 43 2D key-points K_{2D} (obtained from projection), a 25 part part-segmentation map P, and silhouette map S.

4.1. Synth

As in SURREAL [43], RGBT-Dog can generate synthetic images and the corresponding paired labels: We can explore the shape and texture PCA spaces by randomly

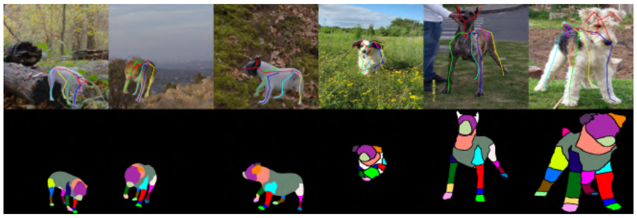

Figure 3. Examples from our Synth dataset (Columns 1–3) and PGT (Columns 4–6). (Top) Images with 2D keypoints overlaid. (Bottom) Part-segmentation maps. White-colored regions indicate *unknown* labels.

sampling shape β and texture γ vectors, respectively to generate new shape and texture. Root translation, t, and rotation, r, are sampled from pre-defined ranges. The pose parameter, θ, is sampled from our motion capture dataset. These are passed to Q (Eq. (5)) and Π (Eq. (6)) to generate images and data labels.

Figure 3 shows examples of our fully synthetic data which we refer to as **Synth**. Using this approach we generated 50,000 instances of synthetic images and the associated data labels for training.

4.2. Pseudo Ground Truth

We can also fit RGBT-Dog to real images to produce the data labels in Eq. (6). We use the real dog keypoints provided by Biggs *et al.* [5]: whose keypoints are a subset of those in RGBT-Dog. Our fitting process is similar to SMPLify-X [36]; we employ a series of priors and supervised losses which are used to optimise a set of parametric model parameters which best allow the model to fit the keypoints provided by Biggs *et al.* [5]. These parameters are then passed to Eqs. (5) and (6) to generate the synthetic data labels for real images. By fitting our model to the images from [5], we generated a training set of 10,309 images with paired synthetic data labels, and test and validation sets each of size 500. We removed some instances of the original 12,000 fits due to impossible fits leaving 11,309. To differentiate this dataset of real images with synthetic data labels from the fully synthetic dataset above, we refer to this dataset as the pseudo ground truth (**PGT**) dataset, as it contains a combination of real images and synthetic labels. Examples of our **PGT** data for keypoint estimation and part-segmentation can be found in Fig. 3 with additional examples in the supplementary material. Note: In the generation of part-segmentation map there is an unknown label (white) which is used to prevent accidental mislabelling.

For fitting RGBT-Dog, we use the real dog keypoints provided by Biggs *et al.* [5]: their keypoints are a subset of those in RGBT-Dog. Our fitting process is similar to SMPLify-X [36]: We employ a loss term between the visible points K_{2D-gt} provided by [5] and the matching points K_{2D} predicted from the fit: $E_J(K_{2D-gt}, K_{2D}) = \rho(K_{2D-gt} - K_{2D})$ with ρ being the German-McClure penalty function: $\rho(x) = \frac{x^2/2}{1+x^2}$. As Biggs *et al.* provides silhouette maps [5], we also employ an auxiliary silhouette loss $E_S = IoU^{-1}$ defined as the inverse of the intersection over union (IoU) between the ground truth map S_{gt} and the corresponding prediction S. Inspired by [48], our fitting also employs an RGB loss E_{RGB} calculated as the perceptual distance [46] between the rendered dog $I_{rgb-ren}$ and the dog in the original image I_{rgb}: We used the silhouette map to remove the background. As in [36], we employ regularisers for our model parameters: $E_\beta(\beta) = \|\beta\|^2$ for shape and $E_\theta(\theta) = \|\theta\|^2$ for pose. Similarly, we also employ L2 regularisation for our texture parameter γ: $E_\gamma(\gamma) = \|\gamma\|^2$. As with SMPLify-X, we employ an angle prior $E_\alpha(\theta) = \sum_{i\in(\text{arm,leg})} \exp(\theta_i)$ to penalise extreme bends of the arms and legs. An interpenetration penalty E_{IP} [36] prevents body parts in the mesh from penetrating the others. To employ our canine pose prior (Sec. 3.3) for fitting optimisation (Eq. (7)) we do not optimize for θ directly (e.g. $\|\theta\|^2$) but rather optimise the parameters of the 40 dimensional latent space using quadratic penalty on said latent space Z which we transform back into pose joint angles θ using the decoder, thus producing a pose. This is the same method used by [36] whose prior inspired our own. This fitting process is summarised by an objective function which we seek to minimise:

$$\begin{aligned} E(\beta,\gamma,\theta,r,t) = {} & \lambda_{RGB}E_{RGB}(I_{rgb}, I_{rgb-ren}) \\ & + \lambda_J E_J(K_{2D-gt}, K_{2D}) + \lambda_\theta E_\theta(\theta) \\ & + \lambda_{IP}E_{IP}(\beta,\theta) + \lambda_S E_S(S_{gt}, S) \\ & + \lambda_\beta E_\beta(\beta) + \lambda_\gamma E_\gamma(\gamma) + \lambda_\alpha E_\alpha(\theta). \quad (7) \end{aligned}$$

For fitting RGBT-Dog to the images and labels of Biggs *et al.* [5] we initially apply high regularisation of pose, shape, 2D key-point loss, and silhouette, with large λ_θ, λ_α, λ_β, λ_J and λ_S values and gradually weaken it as convergence occurs. During this process, for λ_{IP}, λ_S and λ_γ, weak regularisation is initially exercised and it is strengthened to refine the fit once a sufficient pose convergence is achieved. Any of these optimisation terms can also be turned off if specified. Initially for the RGB and texture losses, we would also apply weak regularisation that is strengthened once a sufficient fit has been reached. However, we found that doing so did not aid fitting so we set the terms λ_{RGB} and λ_γ to zero; i.e. we did not fit RGBT-Dog by optimising for these losses.

5. Experiments

Our expressive parametric model can easily generate datasets that provide a rich coverage of motions, geometry and shapes and it can be used in a variety of applications. To demonstrate the real-world utility of such datasets, we present a set of experiments for 2D dog pose estimation and part-segmentation; the latter of which is relatively unexplored with respect to animals. We use a two-stack stacked hourglass [33] for conducting pose estimation and part-segmentation, using the mean squared error and softmax cross entropy losses respectively for supervised learning. For part-segmentation, we assigned a weight of zero to the 'unknown' label in the loss function to prevent our model from learning the unknown label.

For both tasks, across all datasets we trained for five epochs with a batch size of ten and a learning rate of 0.001. For data augmentation, we employed random horizontal and vertical flipping, Gaussian blur, hue saturation and random noise using ImgAug [17].

Regarding part-segmentation, the **Synth** maps possess labels for the eyes whereas the maps for the **PGT** do not. In order to enforce parity with respect to the number of data labels we fold the eye labels of the **Synth** dataset into the head label during the data loading (i.e. these three parts are given the same head label).

We use three datasets: our **Synth** and **PGT** datasets , and a combination of the two which we refer to as the **Mixed** dataset (**Synth + PGT**). For testing these supervised models, the five hundred image **PGT** test dataset were used. We investigated the tasks of canine pose estimation and part-segmentation for all three of these datasets. It should be noted that our goal is to illustrate the utility of our approach for solving complex vision tasks on animals, and the relative performance of the datasets produced by RGBT-Dog, rather than achieving state-of-the-art results. We apply cropping as is standard in body analysis tasks. Additional details regarding training parameters and conditions can be found in the supplementary material.

6. Results

Keypoint estimation. Table 2 shows results for 2D keypoint estimation. We use the percentage correct keypoints (PCK), where a prediction is considered correct if the distance between the predicted and ground-truth joint is less than 20% of the distance between the head and the average of the finger keypoints. We present the PCK results averaged over parent body parts in Tab. 2. Table 2 shows that the performance for **Synth** lags behind the other two datasets showing the effect of the domain gap between the real and synthetic images, as we expected. This gap is in part due to the lack of sitting/resting poses in the **Synth** dataset which are present in **PGT** and the general gap in realism of the **Synth** dogs compared to those of real images. Such a performance gap could potentially be tackled through methods such as adversarial domain adaptation. As is to be expected, the points for the limbs and tail under perform those of the head, neck and torso/spine. This is to be expected given the high degree of mobility and frequent obscuration of such body parts, especially the upper limbs. Full results and additional analysis can be found in the supplementary material. Examples of predicted keypoints are shown in Fig. 4 and additional examples can be found in the supplementary material.

Table 2. PCK results on the PGT test set.

	PGT	Mix	Synth
Spine	59.13	62.67	33.60
L.Arm	56.77	59.37	24.60
R.Arm	55.00	57.07	25.87
Neck	76.80	79.75	38.35
Head	69.33	67.13	29.87
L.Leg	37.20	40.68	16.52
R.Leg	40.28	40.56	16.28
Tail	39.48	41.92	17.47
Average	52.89	54.65	24.19

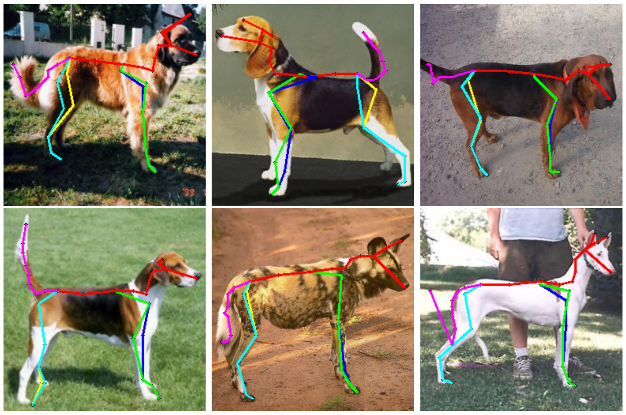

Figure 4. Predicted keypoints from hourglass trained with mixed data on images from [5]. One can see that accurately predicting keypoints for tail and ears is challenging.

We also present some examples of the results of our **Mixed** data trained model when applied to non canine animals (sourced from [45]) in Fig. 5 with impressive accuracy for non tail joints. To some degree this is to be expected given certain poses (e.g. standing) are fairly uniform across quadrupeds), though given different body shapes, predictions for head and tail keypoints suffer in comparison to body and limb joints. Additional examples and quantitative results can be found in the supplementary material.

Part segmentation. Table 3 summarises the results of our part-segmentation experiments. We use intersection over union (IoU) to evaluate part-segmentation results ignoring the 'unknown' label during evaluation. The body parts are averaged into parent groups for ease of presentation and the full results can be found in the supplementary material. As with our pose estimation results, Table 3 shows a significant gap between the **Synth** trained model and the **Mixed/PGT**

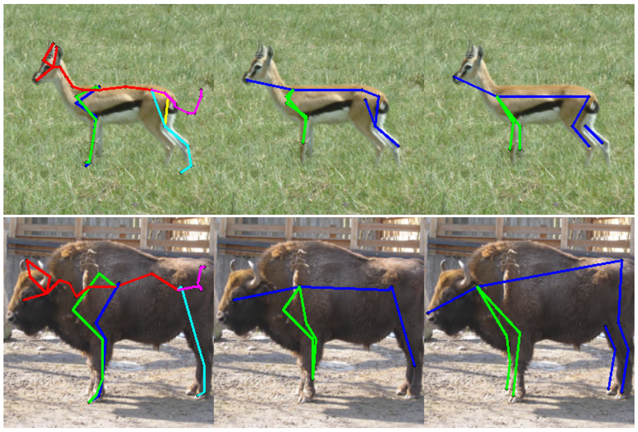

Figure 5. Images from AP10K [45] with keypoints overlaid. From left to right: Predicted RGBT-Dog joints, RGBT-Dog joints that appear in AP10K, ground truth AP10k joints.

trained models as a result of the domain gap between real and synthetic images. In addition to visual realism, the variety of textures in the **Synth** data should also be considered when it comes to closing the domain gap; perhaps with a larger, more realistic PCA texture space, this gap in results could be closed. Domain adaptation could also be employed to close the performance gap. We also note that as in our pose estimation results (Tab. 2), results for the limbs and tail fall below those of the torso, and head across the datasets. The limbs as noted above are frequently obscured and easy to mistake for one another.

Figure 6 illustrates some example results for part-segmentation. We can see that our **Mixed** trained model learns to correct the unknown label in the **PGT** data as shown in columns 3 and 5 where the right back leg and right ear have the correct labels applied. We should also note that the quantitative results in Tab. 3 do not take such ”corrections” into account.

In Fig. 7 we provide some examples of results when our model trained on **Mixed** data is applied to animals other than canines (sourced from [4, 23]). Despite the considerable variation in appearances across different species, we have achieved remarkable results. However, it’s important to note that the effectiveness of our approach varies depending on the animal species. For instance, species like lizards and seals exhibit significantly distinct body shapes and features, which limit the applicability of our method to certain labels pertaining to these animals. Additional examples can be found in the supplementary material.

Model fitting. Here, we provide some metrics for comparing the performance of RGBT-Dog to other parametric canine models when fitting to the images of Biggs *et al.* [5]. We use the same PCK metric as in previous work [5, 39] on the same images for the same keypoints. We also include results for the predictions from our stacked hourglass model trained on **Mixed** data as a point of comparison. However, we should note that RGBT-Dog does not possess chin/mouth keypoints and thus results for the face for RGBT-Dog are calculated from face and nose only. The results are given in Tab. 4 where we can see that RGBT-Dog significantly outperform the previous state of the art with respect to recovering keypoints. We also provide results from directly fitting SMALR to the keypoints of [5]. Our RGBT-Dog model outperforms SMALR on average and across most joint groups though falls behind with respect to the ear joints. We theorise that this is a result of our

Table 3. IoU scores for part-segmentation.

	PGT	Mix	Synth
Background	85.66	86.36	73.58
Torso	47.75	50.03	23.67
L.Arm	22.96	25.68	2.80
R.Arm	25.25	26.83	5.09
L.Leg	23.36	22.86	2.58
R.Leg	23.92	20.87	4.44
Tail	29.70	30.40	3.93
Whole Head	35.03	37.84	8.53
Mean	30.01	30.87	8.48

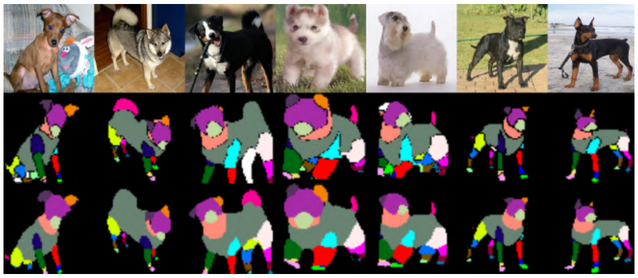

Figure 6. From top to bottom: Images from [20], PGT part segmentation maps, part-segmentation maps predicted by stacked hourglass model trained on mixed data.

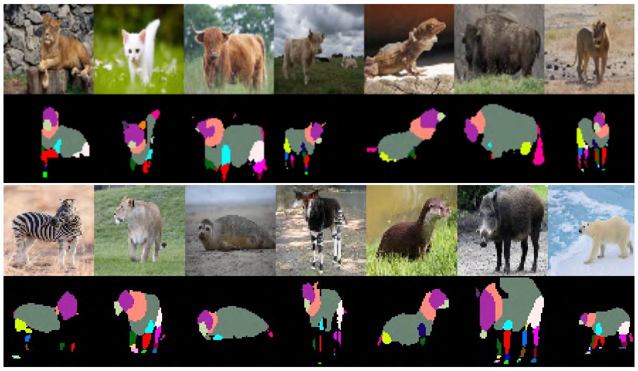

Figure 7. Predicted part-segmentation maps for non-dogs.

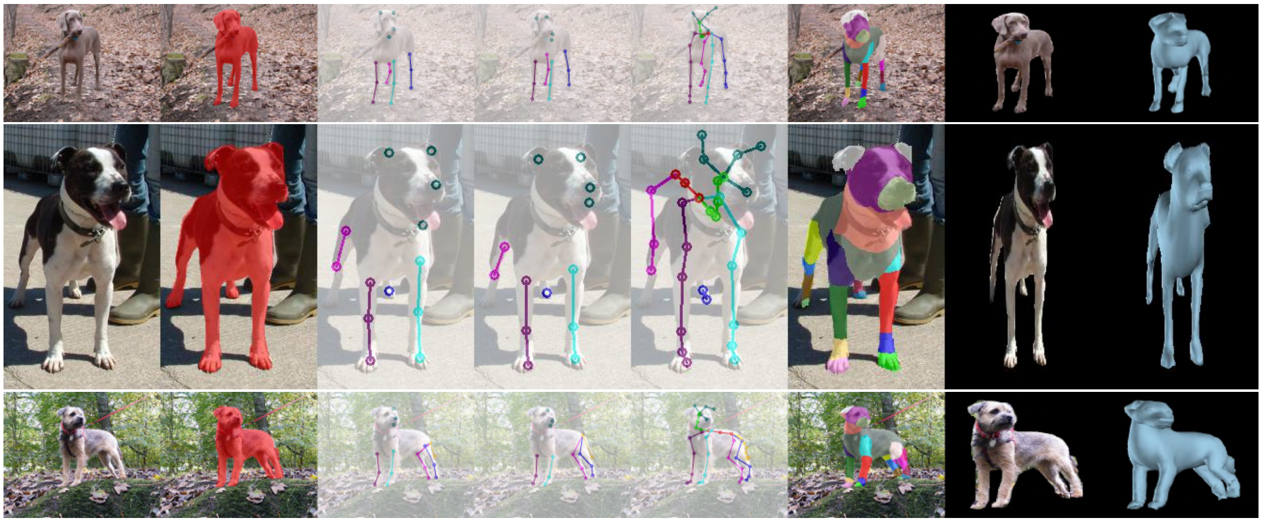

Figure 8. Left to right: Input image, silhouette from [6], keypoints from [6], keypoints from SMALR [50], keypoints from RGBT-Dog, and part-segmentation map from RGBT-Dog, canine image with background masked out, and Canine with RGBT-Dog mesh overlaid.

variational pose prior and the ear end keypoints; most of the dogs that comprised our prior possessed 'short' ears. As a result our model struggles due to the lack of relevant prior information to leverage with respect to the ear end keypoints leading to poorer performance for these keypoints.

Nonetheless, the results of employing our variational pose prior are clear: while our overall improvements for Ears and Face are minimal (and must be taken with a grain of salt for the latter given RGBT-Dogs lack of chin keypoint), our results for Tail in particular show a notable improvement though it should be noted that these tail results are calculated on the two tail keypoints of [5]. By comparison our trained hourglass model shows far less notable results, severely under-performing the fitting methods. This is to be expected given the simplistic nature of our prediction model compared to the fitting methods for the parametric models. Some examples of the fitting results for RGBT-Dog are shown in Fig. 8 and additional examples can be found in the supplementary material.

7. Conclusion

We have introduced RGBT-Dog, a new 3D parametric body model for canines, complemented by an extensive dataset of canine motion capture data, the largest available resource for the community. We used this motion capture data to construct the first variational pose prior specific for canines and indeed animals in general. By employing this prior, we successfully aligned RGBT-Dog with real dog images, generating comprehensive and dense multi-task labels for various body analysis tasks. Our experiments demonstrated that training models using this data, allows for exceptional generalization to other animal species without the need for additional training. These findings highlight the efficacy and versatility of RGBT-Dog as a powerful tool for advancing animal-related research and analysis.

We will release our parametric model, data labels, motion capture data and variational canine pose prior to the community. Given the focus of this work on animal body analysis we do not foresee any negative societal impact of the material in its current iteration.

Table 4. Comparison of canine keypoint recovery methods. PCK@0.15 calculated using the method of [39] on the test data of [5]. Metrics for 3D-M, CGAS, WLDO and BARC are reproduced from BARC [39]. Results for SMALR are obtained from fitting the model using the method of [50]. Results from our stacked hourglass model trained on Mixed data are shown in the final row. *(RGBT-Dog, and by extension our Mixed predictions, lack mouth/chin keypoints so our results for the Face points are simply calculated on the nose and face keypoint).

Method	Avg	Legs	Tail	Ears	Face
3D-M [48]	69.7	68.3	68.0	57.8	93.7
CGAS [6]	28.6	30.7	34.5	25.9	24.1
WLDO [5]	78.8	76.4	63.9	78.1	92.1
BARC [39]	74.2	82.8	63.3	83.3	91.3
SMALR [50]	94.1	95.5	93.4	92.9	90.8
RGBT-Dog*	94.9	96.7	94.5	90.5	98.0
Mixed*	53.5	51.7	61.6	51.9	75.9

Acknowledgements

KIK was supported by the National Research Foundation of Korea (NRF) grant (No. 2021R1A2C2012195).

References

[1] Carnegie Mellon University - CMU Graphics Lab - motion capture library. 4

[2] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1455, Boston, MA, USA, June 2015. IEEE. 4

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, James Davis, and Jim Rodgers. SCAPE: Shape Completion and Animation of People. *ACM Trans. Graph*, 24:408–416, 2005. 2

[4] Sourav Banerjee. Animal Image Dataset (90 Different Animals) (version 3), 2021. 7

[5] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3D animal reconstruction with expectation maximization in the loop. In *European Conference on Computer Vision (ECCV)*, pages 9498–9507, July 2020. arXiv: 2007.11110. 1, 2, 3, 5, 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)*, Nov. 2018. arXiv: 1811.05804. 1, 2, 8

[7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. *arXiv:1607.08128 [cs]*, July 2016. arXiv: 1607.08128. 2

[8] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9497–9506, Aug. 2019. arXiv: 1908.05806. 1, 2

[9] Kai Chen, Wanli Ouyang, Chen Change Loy, Dahua Lin, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, and Jianping Shi. Hybrid task cascade for instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4969–4978, Long Beach, CA, USA, June 2019. IEEE. 1

[10] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3D pose estimation. In *Fourth International Conference on 3D Vision (3DV)*, pages 479–488, Stanford, CA, USA, Oct. 2016. IEEE. 2

[11] Luca Del Pero, Susanna Ricco, Rahul Sukthankar, and Vittorio Ferrari. Articulated motion discovery using pairs of trajectories. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2151–2160, Boston, MA, USA, June 2015. IEEE. 1

[12] Luca Del Pero, Susanna Ricco, Rahul Sukthankar, and Vittorio Ferrari. Behavior discovery and alignment of articulated object classes from unstructured video. *International Journal of Computer Vision*, 121(2):303–325, Jan. 2017. 1

[13] Abassin Sourou Fangbemi, Yi Fei Lu, Mao Yuan Xu, Xiao Wu Luo, Alexis Rolland, and Chedy Raissi. ZooBuilder: 2D and 3D pose estimation for quadrupeds using synthetic data. *arXiv:2009.05389 [cs]*, Sept. 2020. arXiv: 2009.05389. 2

[14] Swarnendu Ghosh, N. Das, Ishita Das, and U. Maulik. Understanding deep learning techniques for image segmentation. *ACM Comput. Surv.*, 2019. 1

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. arXiv: 1703.06870. 1

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, July 2014. 4

[17] Alexander Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gabor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vellentin, Semen Zhydenko, Killian Pfeiffer, Ben Cook, Ismael Fernandez, Francois-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Rapheal Meudec, Mathias Laporte, et al. imgaug, 2020. 6

[18] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, Dec. 2017. arXiv: 1712.06584. 2

[19] Sinead Kearney, Wenbin Li, Martin Parsons, Kwang In Kim, and Darren Cosker. RGBD-Dog: Predicting canine pose from RGBD sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. arXiv: 2004.07788. 1, 2, 3, 4

[20] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford Dogs. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, June 2011. 7

[21] Diederik P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, {ICLR} 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 4

[22] Ivan Krasin, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andres Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. 1

[23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images Dataset V4. *International Journal of Computer Vision*, 128(7):1956–1981, July 2020. 1, 7

[24] Jessy Lauer, Mu Zhou, Shaokai Ye, William Menegas, Tanmay Nath, Mohammed Mostafizur Rahman, Valentina Di Santo, Daniel Soberanes, Guoping Feng, Venkatesh N. Murthy, George Lauder, Catherine Dulac, Mackenzie W. Mathis, and Alexander Mathis. Multi-animal pose estimation and tracking with DeepLabCut. Technical report, Apr. 2021. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article. 1

[25] Chen Li and Gim Hee Lee. From synthetic to real: unsupervised domain adaptation for animal pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1482–1491, 2021. arXiv: 2103.14843. 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. arXiv: 1405.0312. 1

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):1–16, Oct. 2015. 2, 3

[28] Meysam Madadi, Hugo Bertiche, and Sergio Escalera. SMPLR: Deep SMPL reverse for 3D human pose and shape recovery. *Pattern Recognition*, 106, 2020. arXiv: 1812.10766. 2

[29] Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W. Mathis. Pretraining boosts out-of-domain robustness for pose estimation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1858–1867, Waikoloa, HI, USA, Jan. 2021. IEEE. 1

[30] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, Sept. 2018. 1

[31] Jiteng Mu, Weichao Qiu, Gregory Hager, and Alan Yuille. Learning from synthetic animals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12383–12392, 2020. 2

[32] Tanmay Nath, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie Weygandt Mathis. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature Protocols*, 14(7):2152–2176, July 2019. 1

[33] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 483–499, Cham, 2016. Springer International Publishing. 6

[34] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. STAR: Sparse trained articulated human body regressor. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision (ECCV)*, volume 12351, pages 598–613. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science. 2

[35] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. STAR: Sparse trained articulated human body regressor. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12351, pages 598–613. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science. 3

[36] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A A Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D Hands, face, and body from a single image. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3, 4, 5

[37] Pedro O. Pinheiro, Ronan Collobert, and Piotr Dollar. Learning to segment object candidates. *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, 2:1990–1998, Sept. 2015. arXiv: 1506.06204. 1

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, May 2015. arXiv: 1505.04597. 1

[39] Nadine Rüegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. BARC: Learning To regress 3D dog shape from images by exploiting breed information. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3876–3884, 2022. 2, 7, 8

[40] Akash Sengupta. Synthetic Training for Accurate 3D Human Pose and Shape Estimation in the Wild. In *British Machine Vision Conference (BMVC)*, 2020. 2

[41] Moira Shooter, Charles Malleson, and Adrian Hilton. SyDog: A synthetic dog dataset for improved 2D pose estimation. *arXiv:2108.00249 [cs]*, July 2021. arXiv: 2108.00249. 2

[42] Daniel Sánchez, Marc Oliu, Meysam Madadi, Xavier Baró, and Sergio Escalera. Multi-task human analysis in still images: 2D/3D pose, depth map, and multi-part segmentation. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, May 2019. 2

[43] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, July 2017. arXiv: 1701.01370. 2, 3, 4

[44] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3661–3671, Montreal, QC, Canada, Oct. 2021. IEEE. 2

[45] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. AP-10K: A Benchmark for Animal Pose Estimation in the Wild. *Thirty-fifth Conference on Neural Infor-*

mation Processing Systems Datasets and Benchmarks Track, 2021. arXiv: 2108.12617. 1, 6, 7

[46] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *arXiv:1801.03924 [cs]*, Apr. 2018. arXiv: 1801.03924. 5

[47] Xinming Zhang, Xiaobin Zhu, Xiao-Yu Zhang, Naiguang Zhang, Peng Li, and Lei Wang. SegGAN: Semantic Segmentation with Generative Adversarial Network. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pages 1–5, Xi'an, Sept. 2018. IEEE. 1

[48] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D Menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, pages 5524–5532. IEEE, 2017. arXiv: 1611.07700. 2, 3, 5, 8

[49] Silvia Zuffi, Angjoo Kanazawa, Tanja Berger-Wolf, and Michael J Black. Three-D Safari: Learning to estimate zebra pose, shape, and texture from images "In the Wild". *International Conference on Computer Vision*, Oct. 2019. 1, 2

[50] Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5524–5532, 2018. 2, 3, 8