Who left the dogs out? Supplementary Material

Benjamin Biggs[1], Oliver Boyne[1], James Charles[1],
Andrew Fitzgibbon[2], and Roberto Cipolla[1]

[1] Department of Engineering, University of Cambridge, Cambridge, UK
{bjb56,ob312,jjc75,rc10001}@cam.ac.uk
[2] Microsoft, Cambridge, UK awf@microsoft.com

1 Dataset curation

In this section, we describe our process for obtaining keypoint and segmentation annotations for the Stanford Dog Dataset [1]. We submit the entire set of 20,580 dog images to the Amazon Mechnical Turk crowdsourcing platform to obtain a set of 20 keypoint and segmentation masks. We overlay 1 bounding box, provided with the original dataset, on the submitted images to identify the specific dog for the annotators to label. Each image was sent to 3 independent annotators for collecting keypoints and segmentation masks.

Keypoints. To identify keypoints, workers were given a list of 20 keypoints to click: 2 per tail, 3 per leg, 2 per ear, nose and jaw. They were additionally asked to provide a visibility flag per point.

For each keypoint, we process the three clicks to yield a reliable coordinate. From the 3 clicks, we discard clicks that are further than a set tolerance from the mean. If at least 2 clicks remain, we take the mean coordinate as the accepted keypoint position. Otherwise, the point has not been reliably identified between workers, so we set the keypoint as invisible. As described in the main paper, we remove images from train and test splits which have fewer than 8 visible keypoints.

Segmentation. For each image, each worker $w \in \{w_1, w_2, w_3\}$ submits a binary segmentation mask $\mathbf{A}^w \in \mathbb{R}^{H \times W}$. We request a re-labelling for any submissions which fail simple criteria, such as if the highlighted area is below a threshold number of pixels.

For each image, we generate the most likely segmentation by comparing submissions across workers. For any two workers w, w' we compute a correlation coefficient:

$$c_{w,w'} = \frac{\sum_i \sum_j \left[\mathbf{A}^w \odot \mathbf{A}^{w'}\right]_{i,j}}{\max\limits_{p=\{w,w'\}} \sum_i \sum_j \mathbf{A}^p_{i,j}} \tag{1}$$

Where $\odot$ denotes the element-wise product of the matrices. We remove a worker's segmentation A^w if all correlation coefficients $c_{w,w'}$ are below a set threshold. The final binary mask is computed from the remaining submissions:

$$\hat{A}_{i,j} = \begin{cases} 1, & \text{if } \sum_w A^w_{i,j} > 1 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

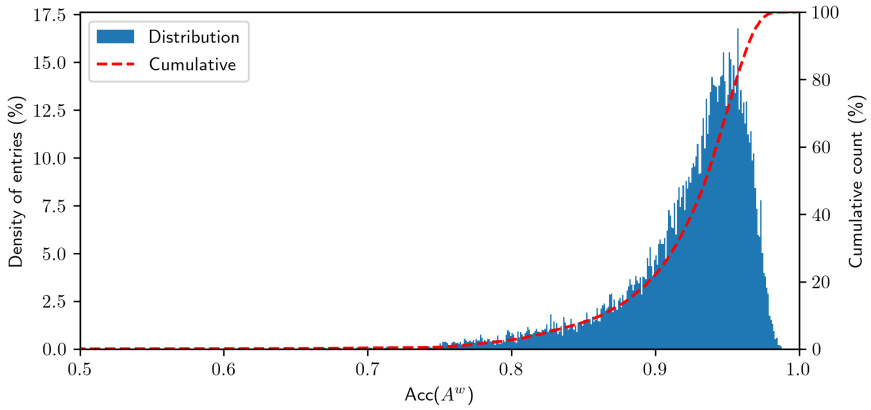

Fig. 1. Accuracy distribution of all submitted dog segmentations across the entire Stanford Dog Dataset.

We can also define the accuracy of a worker's segmentation, as the largest of their correlation coefficients: $\text{Acc}(A^w) = \max_{w' \neq w}\{c_{w,w'}\}$. Figure 1 shows the set of segmentation annotation accuracies over the entire labelled dataset.

2 Fitting SMBLD to 3D animation data

Another method for improving the generalizability of the SMAL model is to improve the 3D shape prior. Such priors are typically used to ensure shape deformation remain within a realistic and anatomically plausible range. Due to the limited diversity of scans used to build the SMAL model, while the shape prior does enforce realism among deformations, it does not allow for a wide enough range to cover the set of dogs in our dataset.

We improve the quality of the prior (and learn a prior over our new scale parameters) by fitting to a set of 13 artist-designed 3D dog meshes, designed for animation use, which are more varied than the original set. We apply an energy minimization scheme which aligns the SMAL vertices to each scan, under smoothing regularizers.

Recall that SMBLD is adapted from the SMAL [3] deformable animal mesh, by including limb scaling parameters. We learn a prior by fitting our SMBLD model, which comprises parameters for pose θ and shape β (the latter of which includes our scaling parameters κ).

Note that fitting SMBLD to 3D scans is significantly easier than to 2D images, since the complete 3D information of the target mesh is available. In addition, our target meshes are not particularly detailed and are already aligned in

T-pose, so we avoid need for a complex alignment technique as discussed in, for example SMPL [2] or SMAL [3].

We run an energy minimization process to align the SMBLD mesh to the 3D scans, subject to some smoothing regularizers. We minimize the following energy formulation:

$$E_{\text{opt}} = E_{\text{chamfer}} + E_{\text{laplacian}} + E_{\text{edge}} + E_{\text{normal}} \tag{3}$$

where each of these terms has a scalar weight λ. We set $\lambda_{\text{chamfer}} = \lambda_{\text{edge}} = 1.0$, $\lambda_{\text{normal}} = 0.01$ and $\lambda_{\text{laplacian}} = 0.1$. We run the optimization using SGD, learning rate 10^{-4} for 1000 iterations.

Chamfer energy. A measure of the average distance between vertices of the SMBLD mesh $V = F_v(\theta, \beta)$, and the target mesh vertices V', when p vertices v_i, v'_j are sampled from each mesh respectively:

$$E_{\text{chamfer}}(V, V') = \frac{1}{p} \sum_{i=1}^{p} \min_{j}^{p} \left| v_i - v'_j \right| \tag{4}$$

Uniform laplacian energy. A measure of the mesh smoothness.

Edge energy. This energy is equal to the average edge length across the mesh, and is used to encourage uniform distribution of vertices.

Normal energy. This energy promotes consistency between adjacent faces. It is a measure of the average normal consistency between adjacent faces. For two faces with normals $\mathbf{n_0}$ and $\mathbf{n_1}$, the normal consistency is $1 - \frac{\mathbf{n_0} \cdot \mathbf{n_1}}{|\mathbf{n_0}||\mathbf{n_1}|}$.

At the end of this process, we have a collection of fits $|(\theta, \beta)|_{\{i=1,\ldots 13\}}$ from which we can learn our unimodal pose and shape priors. As discussed, we evenutally use this unimodal shape prior to initialize our mixture shape prior, which is tuned with the expectation-maximization step in the training loop.

3 Training procedure

Recall that the training objective for our end-to-end system for predicting SMBLD parameters consistent with a monocular dog input image is given by:

$$L_{\text{opt}} = L_{\text{joints}} + L_{\text{sil}} + L_{\text{pose}} + L_{\text{shape}} + L_{\text{mixture}} \tag{5}$$

As described in the paper, each loss term is weighted with a scalar λ and we train our method in two stages:

Stage 1. We set $\lambda_{\text{joints}} = 10.0, \lambda_{\text{pose}} = 1.0, \lambda_{\text{shape}} = 1.0, \lambda_{\text{sil}} = 0.0, \lambda_{\text{mixture}} = 0.0$. We train this stage for 250 epochs, using the Adam optimizer, with learning rate set to 10^{-4}.

Stage 2. In this stage, we introduce the silhouette loss to encourage a shape alignment between the projected model silhouette and the ground truth annotation. We set $\lambda_{\text{joints}} = 10.0, \lambda_{\text{pose}} = 0.5, \lambda_{\text{shape}} = 0.0, \lambda_{\text{sil}} = 100.0, \lambda_{\text{mixture}} = 0.1$. We train this stage for 150 epochs and run the described EM update step every $K = 15$ epochs. We selected to use $M = 10$ clusters based on a grid search over $M = 1, 5, 10, 25$ and comparing IoU. We again use the Adam optimizer, and set the learning rate to 10^{-5}.

References

1. Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L.: Novel dataset for fine-grained image categorization. In: First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO (June 2011)
2. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 248 (2015)
3. Zuffi, S., Kanazawa, A., Jacobs, D., Black, M.J.: 3D menagerie: Modeling the 3D shape and pose of animals. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Jul 2017)