Birds of a Feather: Capturing Avian Shape Models from Images **Supplementary Material**

Yufu Wang Nikos Kolotouros Kostas Daniilidis Marc Badger
University of Pennsylvania
{yufu, nkolot, kostas, mbadger}@seas.upenn.edu

This supplementary material provides additional details that are not included in the main text due to space constraint. In Section 1, we provide details about the scaling parameter that augments the template model. In Section 2, we discuss how subdivision may affect our method and downstream applications. Section 3 provides implementation details for single-view model regression, including the model architecture and training procedure. Section 4 presents further exploration of the learned shape space and its relation with the avian phylogeny. Section 5 provides 3D evaluation on the main method. Finally in Section 6, we include more qualitative examples from species-specific captures and single-view regression.

1. Local Scaling Parameter: κ

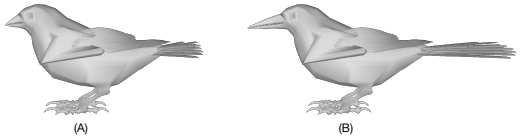

Figure 1: Part Scaling. We augment the template model with two local scaling parameters. (A) is the default, and (B) is with 2x scaled beak and tail.

Our method assumes accurate alignment of the template model to annotated instances. This alignment is more difficult if the new species' beak or tail has very different length than the template's. Similar to [3], we augment the template model with a local scaling parameter, $\kappa \in R^2$, that scales the beak and the tail independently at their local coordinates, and along the longitudinal direction. Uniform scaling alone does not capture high fidelity details, but the scaled beak can be further refined by the deformation steps (main text **Section 3.3** and **3.4**).

2. Model Subdivision Level

Different subdivision levels can be applied to the template articulated mesh. Previous methods demonstrate benefits of using subdivided surface [1, 4]. To achieve faster computation, we use no subdivision (level 0) during the optimization, but subdivide the learned models to create a smoother version after learning. That is, we learn our models at level 0, and propose using them at level 0 or 1. Comparing the non-subdivided and the subdivided mean shape of AVES in Figure 2, we see that subdividing [14] the surface once does not destroy important structures learned at level 0, such as the crest, the beak and the overall body shape, but improves the realism in the parts that are purposefully designed by the artist for subdivision, such as the eyes and the wing elbows.

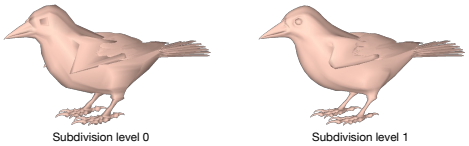

Figure 2: Subdivision Level. We learn our models at subdivision level 0, and create a smoother version by subdividing the surface once. Shown here are the mean shape of AVES at two different subdivision levels.

3. Regressing from an RGB Image

In this part we provide more details about the model we used and the training strategy. Our model is based on the popular HMR architecture [9]. More specifically we use the same ResNet50 [7] backbone and make some small modifications in the decoder. Instead of predicting the 85 SMPL and camera parameters we predict the parameters of the AVES body model together with the camera translation. We follow SPIN [13] and use the 6D representation for rotations [18] and a full perspective camera model. Since there is less variation in the pose of birds compared to humans we use only a single iteration of the iterative regressor. The total loss is

$$L = \lambda_1 * L_{keypoints} + \lambda_2 * L_{silhouettes} + \lambda_3 * L_\beta + \lambda_4 * L_\theta + \lambda_5 * L_\alpha \quad (1)$$

where $L_{keypoints}$ is an L1 loss on the reprojected 2D keypoints, $L_{silhouettes}$ an L2 loss on the reprojected keypoints and L_β, L_θ and L_α L2 regularization losses on the shape, pose and bone length parameters respectively. For the shape parameters, $L_\beta = ||\beta||_2^2$ whereas for the pose and bone length parameters $L_\theta = ||\theta - \hat{\theta}||_2^2$ and $L_\theta = ||\alpha - \hat{\alpha}||_2^2$, The mean θ vector, $\hat{\theta}$, is the same as the one used in [2], whereas the mean bone length vector is a vector of ones.

For the loss weights we used $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 0.001$, $\lambda_4 = 0.05$ and $\lambda_5 = 0.1$.

We used the ground truth boxes to crop the images around each bird and then resized the crops to 256×256. The size of the boxes was rescaled by a factor of 1.1 as in CMR [10]. At training time we used a combination of random augmentations. We randomly rescaled and translated each the ground truth box, applied random rotations and flipping, and also performed color jittering in the RGB image.

We trained the neural network with a batch size of 64 images using the Adam optimizer [12] with a learning rate of 0.0001 and 0.0001 for weight decay. We trained for a total of 500 epochs.

We show additional regression results and comparisons in Figure 7. Some typical failure cases are shown in Figure 8.

4. Shape Space Analysis

In biology, morphometrics uses statistical shape analysis to study the variation and patterns of body shape among species. Changes in shape over evolutionary time reflect both the relatedness among species (i.e. the phylogeny) as well as changes introduced by random mutation or selective pressures. Analyzing shape in the context of the phylogeny can reveal whether species have similar forms because of their relatedness or because of convergent evolution. Analyses can also identify species or groups that have undergone directional selection, arriving at a very different morphology than that of the common ancestor of related species.

4.1. Ancestral state reconstruction

We start with a complete, dated phylogeny [8], which was constructed using all 9,993 living bird species and calibrated in time using fossil data. From this phylogeny, we extract the tree for the subset of the 17 species whose shape we capture (shown in Figure 3). After capturing the shape of each individual, we visualize the shape space coefficients for all individuals using a 2D UMAP embedding (Figure 5a of the main text). We then use `Rphylopars` [6, 5] to reconstruct the ancestral state of the branch points in the phylogeny assuming a Brownian motion model of trait evolution for the two embedding coordinates. We perform an identical process on the UMAP embedding of the learned perceptual features (Figure 5b of the main text).

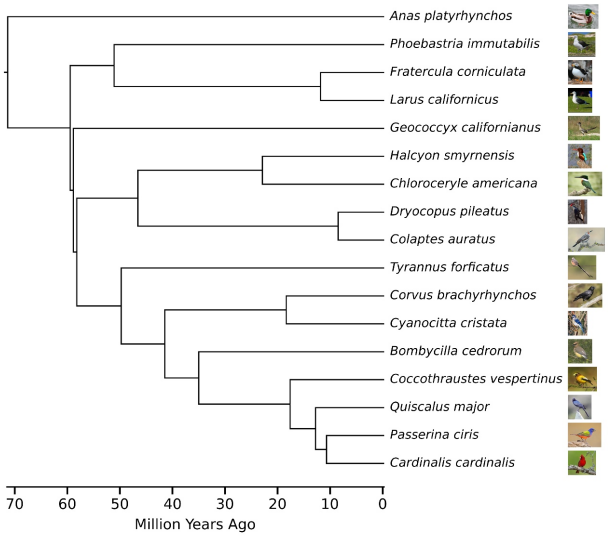

Figure 3: **Phylogeny of captured species.** Extracted from a complete, dated avian phylogeny [8].

Config.	Features	lambda	p-value
dims $= 2$, nn $= 16$			
	AVES Shape PCs	0.97 ± 0.10	0.014
	ResNet50	0.45 ± 0.25	0.29
dims $= 2$, nn $= 75$			
	AVES Shape PCs	1.00 ± 0.01	0.0003
	ResNet50	0.10 ± 0.16	0.81
dims $= 7$, nn $= 16$			
	AVES Shape PCs	0.99 ± 0.02	< 0.0001
	ResNet50	0.18 ± 0.20	0.60
dims $= 7$, nn $= 75$			
	AVES Shape PCs	0.97 ± 0.05	0.0002
	ResNet50	0.34 ± 0.13	0.23

Table 1: **Phylogenetic signal and likelihood ratio tests using 2D and 7D UMAP embeddings.** UMAP parameters include `dims`, the number of embedding dimensions, `nn`, the number of neighbors (lower captures more local structure, higher captures more global structure), and `min_dist`, the minimum allowed distance in the embedding between two samples. We set `min_dist` $= 0.9$ for all experiments. P-values are for likelihood ratio to tests for phylogenetic signal assuming a star phylogeny as the null hypothesis. All values are mean $\pm$ standard deviation across 100 replications.

4.2. Phylogenetic signal of shape and appearance traits

To calculate the phylogenetic signal of shape space coefficients vs. learned perceptual features, we take the first 256 shape coefficients and generate a 7-dimensional UMAP embedding (with parameters `min_dist` $= 0.9$ and `n_neighbors` $= 16$). We extract learned perceptual features from a ResNet50 embedding network, which is trained on CUB using Proxy-Anchor loss and outputs 512 features (trained network weights are provided by [11]). We then generate another 7-dimensional UMAP embedding from these perceptual features with the same UMAP parameters. The 7D shape and perceptual features serve as "traits" from which we calculate Pagel's lambda [15, 16, 17]. Because UMAP is a stochastic algorithm, Table 1 of the main text presents the mean and standard deviation across 100 embedding replications with different random initializations.

For each replication, we also calculate the p-value for a likelihood ratio test comparing the probability of the observed lambda given the avian phylogeny versus a "star" phylogeny where all species' branches originate from a single common ancestor node and lambda is zero (i.e. a tree where all species are equally related). In Table 1 of the main text, we report the mean of these p-values.

We perform the same analysis using various combinations of UMAP parameters and find identical results. We try embedding in either 2 or 7 dimensions and we try two values for the `n_neighbors` parameter, which governs fidelity to local versus global structure (we try 16 and 75, a quarter of the dataset; lower captures more local structure, higher captures more global structure). Results are in Table 1.

5. Evaluating on 3D data

One major obstacle in 3D animal reconstruction is the lack of large scale 3D benchmarks. Usually the accuracy of different methods is measured using 2D metrics such as keypoint or silhouette reprojection errors, which may not reliably reflect the accuracy of the 3D shape reconstruction.

To evaluate our method, we additionally acquire 6 different toucan meshes from the internet to render a synthetic dataset for quantitative 3D evaluation. The 6 toucan meshes are from different artists, with different shapes, poses and realism; they simulate different instances of the same category in our experiment. We create 5 sets of images, with each set consisting of one random view per instance, and average results over the 5 sets. Because we only have 6 samples per set, we avoid obscure frontal views. We render all annotations and assume no occlusions.

We run our method on the rendered dataset. In Figure 4 we can see an example of this process. For evaluation, we rigidly align the reconstruction with the ground truth, and compute 3D keypoint distances and scan-to-mesh distances to the ground truth meshes. Table 2 shows the results of the quantitative evaluation. All numbers are expressed in terms of percentage of each toucan's body length, measured from bill tip to tail tip. We can see that each additional step (part scaling, adding mean per-species offsets, and identity-specific deformation) improves the performance over the previous level.

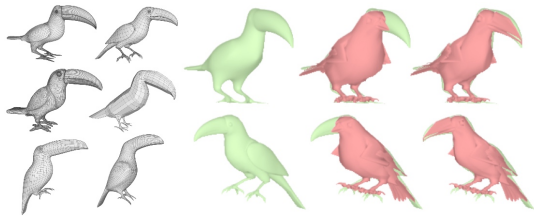

Figure 4: **Toucan experiment.** Gray: toucan mesh data. Color: from left to right are the ground truth rendering, initial alignment, and final reconstruction.

Metrics	alignment	+part scaling	**+dv**	**+dv+V$\beta^{(i)}$**
3D keypoints	9.07%	8.31%	7.59%	**7.37%**
scan-to-mesh	15.33%	6.48%	5.14%	**4.96%**

Table 2: **Quantitative evaluation on toucan data.** Numbers are averaged distances expressed as percentage of the body length (the lower the better).

6. More Examples

We present more reconstruction examples for:

- Species-specific capturing (Figures 5 and 6).
- Regressing AVES parameters from a single RGB image (Figures 7 and 8).

6.1. Results of Species-specific Capturing

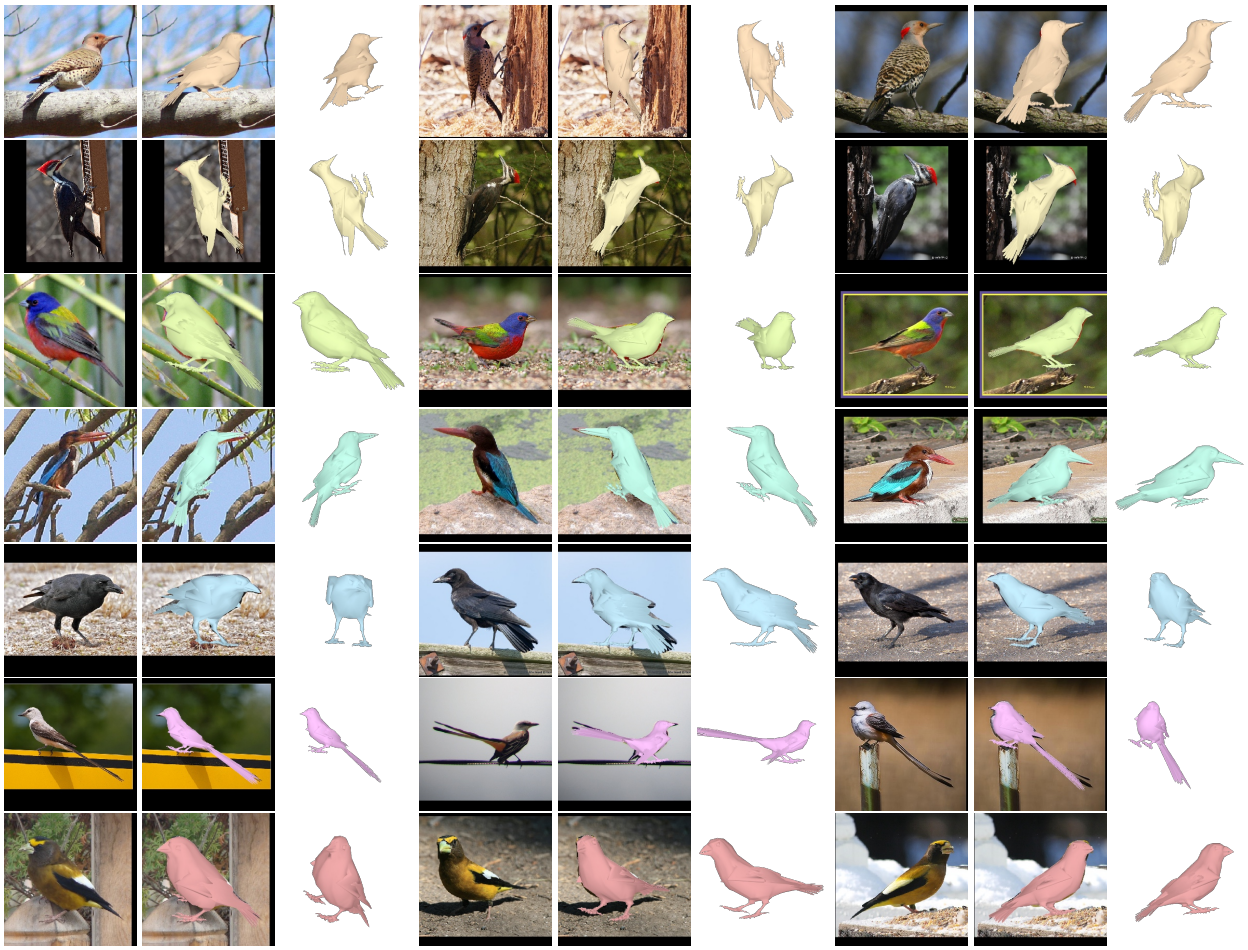

Figure 5: **Examples of species-specific captures** of the 7 species that are not visualized in the main text. Each row depicts reconstructions using a particular species-specific model. Each triplet includes the input image, the reconstructed mesh and the reconstructed mesh from a novel viewpoint.

Figure 6: **Failure cases.** First row are typical examples that fail pose alignment (Sec. 3.2) and are not included in the deformation steps (Sec. 3.3-3.4). The bottom two rows are failure results from the deformation steps. Failure modes include perspective foreshortening, unrealistic deformation, and failing to capture more intricate details.

6.2. Regression using AVES

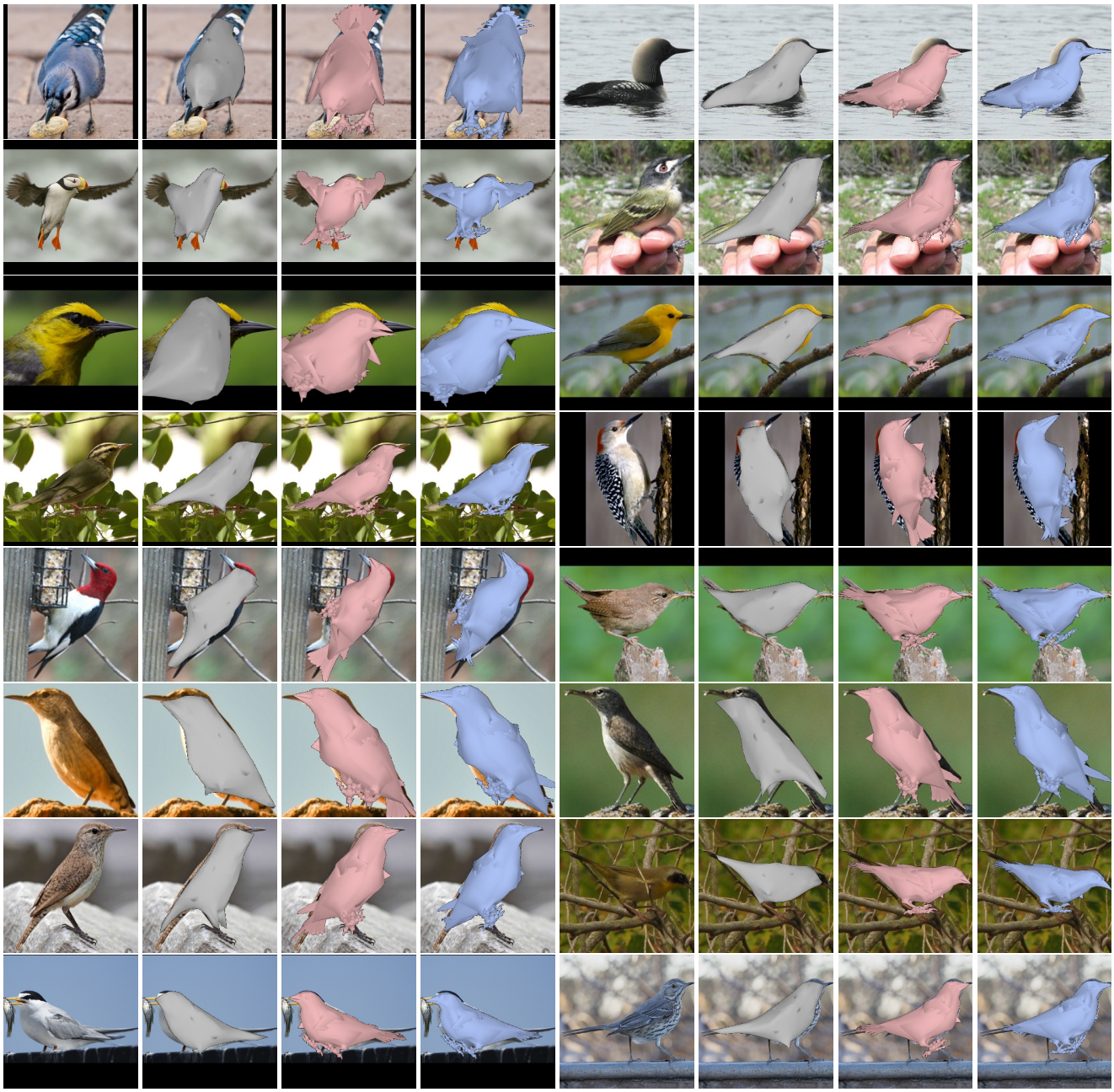

Figure 7: **Qualitative comparison of regression-based methods.** Gray: Reconstruction by CMR [10]. Pink: Baseline (ABM) [2]. Blue: Ours (AVES).

Figure 8: **Failure cases of our regression method.** Typical examples of failure modes are unusual poses, extreme articulation (open wings) and articulation outside the model space (open beak).

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 1

[2] Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd G Pfrommer, Marc F Schmidt, and Kostas Daniilidis. 3d bird reconstruction: a dataset, model, and shape recovery from a single view. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 2, 5

[3] 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*, pages 195–211. Springer, 2020. 1

[4] Thomas J Cashman and Andrew W Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):232–244, 2012. 1

[5] Eric W. Goolsby, Jorn Bruggeman, and Cecile Ane. *Rphylopars: Phylogenetic Comparative Tools for Missing Data and Within-Species Variation*, 2019. R package version 0.2.12. 2

[6] Eric W. Goolsby, Jorn Bruggeman, and Cécile Ané. Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution*, 8(1):22–27, 2017. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[8] W. Jetz, G. H. Thomas, J. B. Joy, K. Hartmann, and A. O. Mooers. The global diversity of birds in space and time. *Nature*, 491(7424):444–448, Nov 2012. 2

[9] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 1

[10] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 2, 5

[11] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR), arXiv preprint arXiv*, volume 1412, 2015. 2

[13] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2261, 2019. 1

[14] Charles Loop. Smooth subdivision surfaces based on triangles. *Master's thesis, University of Utah, Department of Mathematics*, 1987. 1

[15] Mark Pagel. Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, 26(4):331–348, 1997. 3

[16] Mark Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884, Oct 1999. 3

[17] Liam J. Revell. Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution*, 1(4):319–329, 2010. 3

[18] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1