Birds of a Feather: Capturing Avian Shape Models from Images

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

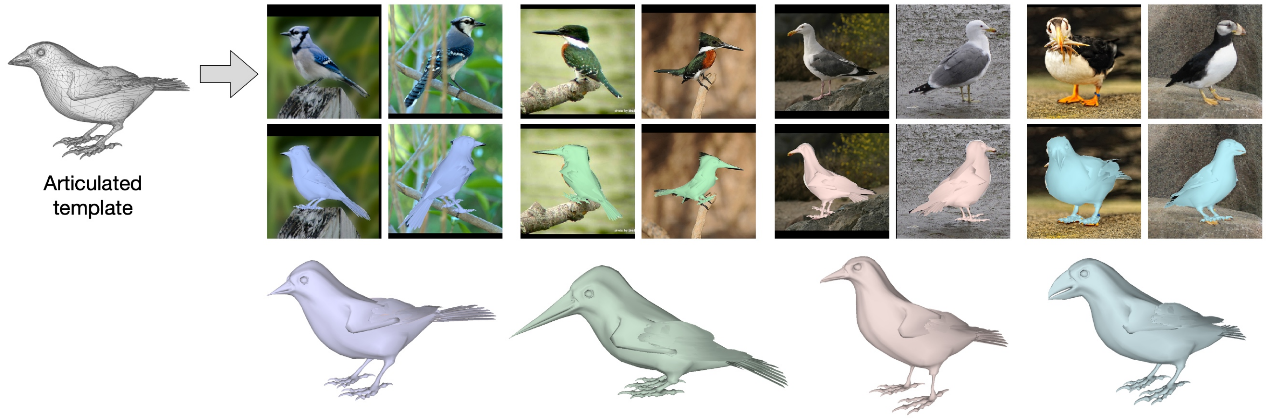

Figure 1: **Capturing shape models from images**. We deform an articulated template to capture species-specific shape models from CUB image collections [50]. The new shape models not only articulate but can also deform according to species-specific shape deformation modes. We combine models from diverse species to learn a multi-species model.

Abstract

Animals are diverse in shape, but building a deformable shape model for a new species is not always possible due to the lack of 3D data. We present a method to capture new species using an articulated template and images of that species. In this work, we focus mainly on birds. Although birds represent almost twice the number of species as mammals, no accurate shape model is available. To capture a novel species, we first fit the articulated template to each training sample. By disentangling pose and shape, we learn a shape space that captures variation both among species and within each species from image evidence. We learn models of multiple species from the CUB dataset, and contribute new species-specific and multi-species shape models that are useful for downstream reconstruction tasks. Using a low-dimensional embedding, we show that our learned 3D shape space better reflects the phylogenetic relationships among birds than learned perceptual features.

1. Introduction

Automated capture of animal shape and motion is a challenging problem with widespread application in agriculture, biomechanics, animal behavior, and neuroscience. Changes in shape can convey an animal's health status and transmit social signals [1, 34]. Recent methods that use articulated mesh models to extract these signals from images [7, 8, 54, 55] are poised to transform these fields. One challenge that prevents wider adoption of this approach, however, is the difficulty of obtaining suitable models for new species. Methods that can automatically capture the shape of new species are highly desirable.

Recent articulated, 3D animal shape models obtain training data from 3D scans of toy animals figurines, or multiple views of the same subject [7, 56, 55]. They produce great quality of reconstruction when the target is represented in the source data. How these articulated and deformable models can be extended to new animal species is still an open problem, particularly due to the lack 3D scans. Animals' non-cooperative nature makes obtaining 3D scans impractical.

Images of the same species, on the other hand, are more readily available for a wider variety of categories. The CUB [50] dataset, for example, provides image collections for various bird species. These monocular collections remain underutilized, however, because current detailed capture methods require a strong deformable 3D shape prior, which is not available for many species.

We propose a method to directly capture articulated shapes from species-specific image collections, when a strong deformable shape prior is not available. We focus our effort on CUB and start with an articulated bird mesh [7] as a generic template model. For a given collection, we first align the mesh to each annotated instance by solving for the pose and camera parameters. We then update the template model through a series of deformations to better fit the silhouettes, resulting in recovery of a new species-specific shape and individual variations within the species.

Closest to our work is SMALR [55], which uses a deformable shape model and a video sequence of the same instance to reconstruct quadrupeds. We relax this assumption and use a collection of different instances. This breaks the same-subject constraints and makes a naive adaption of SMALR infeasible. Solving this "multiple instances" problem allows us to build species-specific morphable shape models directly from images. Our method also starts with a simpler shape prior, using an articulated mesh as the template model.

To handle these challenges, we explicitly model two levels of the shape hierarchy. The first level is the difference between the generic template model and the average shape of a new species; the second level is the variation among individuals within that species.

First, after aligning the articulated mesh to the images for a novel species, we optimize a per-vertex deformation to bridge the difference between the shape of the template model and the shape of a novel species. We call the resulting shape the species mean shape. In the second step, starting from the mean shape, we learn the variation within the collection as a blend shape basis, allowing us to reconstruct each sample as a combination of shape basis vectors on top of the estimated mean. The shape basis also provides a species-specific morphable shape model. Because the articulation is factored out during model alignment, the mean and shape basis are properly captured without the nonlinear effect of pose.

Additionally, from all the per-species models we learn a new parametric shape model, AVES, that is capable of representing multiple avian species. We demonstrate through experiments that AVES captures a meaningful shape space, generalizes to unseen samples, and can be integrated in a deep learning pipeline.

In summary, our contributions are the following:

- We present a new method that recovers detailed, species-specific shape models for a range of bird species using an articulated 3D mesh and images of different instances of each species.
- We provide AVES, a multi-species statistical shape model extracted from reconstructions of 17 bird species. We show that the AVES shape space captures morphological traits that are correlated with the avian phylogeny.
- We show that the AVES model can be used in downstream reconstruction tasks including model fitting and regression, outperforming previous model-based and model-free approaches by a large margin.

2. Related Work

Our goal is to capture articulated 3D animal shapes from species-specific images. Here we focus on approaches that are most relevant to our problem, and review how they have been applied to humans and animals.

Model-based Reconstruction. The reconstruction of non-rigid and articulated objects benefits greatly from a strong prior. A wealth of methods employ parametric models and treat the reconstruction problem as a parameter estimation problem. For human body, such models are learned from thousands of registered 3D scans [5, 6, 23, 32, 38, 41, 52]; SMPL [32] being the most widely used.

For animals where 3D scans are impractical, various adaptations have been proposed. Zuffi *et al*. [56] introduced SMAL, an articulate quadruped model parameterized similar to SMPL and learned from 3D scans of toy figurines. Biggs *et al*. [8] extend the SMAL model to include limb scaling to model dog breeds. Badger *et al*. [7] similarly parametrized an articulated bird mesh and used multi-view data from an aviary to provide pose and shape priors.

These models can be fitted to different modality of sensor data or annotations in an optimization paradigm [10, 21, 51, 56]. Deep learning has made directly regressing model parameters possible [24, 30, 29, 37, 42]. Similarly, these techniques are adapted to regress parameters for articulated animal models when training data is available [7, 8].

Recovery beyond Parametric Models. Building details on top of a parametric model has the advantage that shape and pose information are decoupled and the recovered shape can be easily re-animated. Alldieck *et al*. [3, 4] use video sequences of the same human to optimize a non-rigid deformation on SMPL to reconstruct details in hair, clothing and facial structure. Octopus [2] then obtains comparable results by training a network with synthetic data to predict the deformation using multiple views and test-time optimization-based refinement. ARCH [22] learns to estimate an implicit surface around a rigged body template to recover detailed and animatable avatars from a single image.

Similarly for animals, Zuffi *et al*. [55] non-rigidly deforms SMAL with video sequences to capture detailed and textured quadrupeds. Three-D Safari [54] uses this method to capture 10 realistic zebras to generate synthetic data and train a network to predict zebra pose and shape.

Our approach also captures deviation from a parametric model. Differently, we relax the requirement of having a single subject in video sequence, and instead capture deformations in a "multiple instance" setting.

Model-free Capturing. There is a large amount of works on human shape capture without a parametric model but they tend to be supervision-heavy and do not immediately generalize to animals. Please refer to [19] for a more comprehensive review.

For animals, "model-free" methods focus on lowering the requirement of supervision or prior knowledge. Early methods employ inflation techniques to extract shapes from silhouettes [36, 49]. More recent methods learn end-to-end predictors, such as CMR [25], U-CMR [17] and IMR [48], to produce textured 3D animal mesh from a single image. So far, the outputs of these methods have limited realism and cannot be re-animated.

Learning Shape Models. We aim to capture the shape space of a new species in the form of a morphable basis model. Shapes are classically defined over the geometry that is invariant only to Euclidean similarity transform [14]. Many traditional deformable models treat the same object (e.g. hands) going under articulation to have different shapes [11, 13, 20, 53]. As a result, their basis encode both pose and shape.

More recent methods often see benefits in disentangling pose and shape for articulated objects [6, 31, 32, 45, 56]. Their shape space captures deformation intrinsic to identity, or animal species. Unfortunately, most of these models are learned from registered 3D data that are generally not available for many animal categories. Other methods have also learned deformable models from images, but they either do not disentangle articulation and shape, or only apply to rigid categories [12, 25, 26, 47]. Our approach uses an articulated mesh to explicitly separate pose before learning a meaningful basis shape space directly from annotated images.

3. Approach

We use the recent parametric bird model [7] as a starting point. For each sample in a species' image collection, we first align the model, through articulation and translation, to the annotated silhouette and keypoints. We then fix the alignment and update the shape in order to improve reconstruction on each image. We decompose this update into two steps. In the first step we optimize a per-vertex deformation on the shape template that is shared for every sample. This step transforms the generic template shape to an estimated mean for the new species. Finally, on top of this mean shape, we optimize a set of shape basis vectors and their coefficients for each sample to capture variation within the collection.

3.1. Articulated bird model

The articulated bird model (ABM) [7] is a function $M(\theta, \alpha, \gamma)$ of pose θ, bone length α, and translation γ. Pose $\theta \in \mathbb{R}^{3J}$ is the axis-angle representation of each joint orientation in the kinematic skeleton. The bone parameter $\alpha \in \mathbb{R}^J$ scales the distance between neighbouring joints. It allows body proportion to slightly vary and additionally models non-rigid joint movements that are common in birds, *e.g.* the stretching of a bird's neck. Finally, $\gamma \in \mathbb{R}^3$ applies a translation to the root joint. The function $M(\theta, \alpha, \gamma)$ then articulates a template mesh $\mathbf{v}_{bird} \in \mathbb{R}^{3N}$ through linear blend skinning and returns a 3D mesh.

We aim to align ABM to different species, but accurate alignment is more difficult when the new species has very different beak or tail length. Similar to [8], we augment the model with two local scaling parameters, κ, that scale the length of the beak and tail respectively. The resulting model becomes $M(\theta, \alpha, \gamma, \kappa)$. Though such scaling is not always realistic, it can be refined during the deformation steps.

3.2. Alignment to images

To align the articulated mesh to images, we adopt a regression followed by optimization approach similar to [7, 54]. First we use a regression network that takes 2D keypoint locations as input and predicts model parameters: α, θ, γ. This is then used as initialization for an optimization procedure that refines the alignment.

The regression network consists of two fully connected layers, each followed by batch normalization and max-pooling, and a final layer that predicts all model parameters. Different than [7], we do not include the silhouette as input to predict α; because the bone parameter α mainly captures body proportion changes and this information can be inferred from keypoints alone. We train the network with synthetic keypoints-parameters pairs that are generated by animating the model with its pose priors.

Using the regression result as initialization, we minimize keypoints and silhouette reprojection error and a pose prior regularization with respect to $\Theta^{(i)} = \{\alpha, \theta, \gamma, \kappa\}$ for each image i independently. We omit notation (i) in this step. Specifically, we define the objective as:

$$E(\Theta) = E_{kp}(\Theta) + E_{msk}(\Theta) + E_{prior}(\alpha, \theta) \qquad (1)$$

Keypoint reprojection. The keypoint reprojection term penalizes the distances between annotated keypoint locations and reprojection of the mesh keypoints. Denoting $P_k(\Theta)$ as a function that returns the k^{th} keypoints from the articulated mesh, $\Pi = \Pi(P_k(\Theta))$ as its projection, $\Pi(x)$ as the camera model, and $\mathcal{P}_k$ as the ground truth, the keypoint term can be expressed as

$$E_{kp}(\Theta) = \sum_{\text{kpt } k} \rho(\| \Pi(P_k(\Theta) - \mathcal{P}_k \|_2) \qquad (2)$$

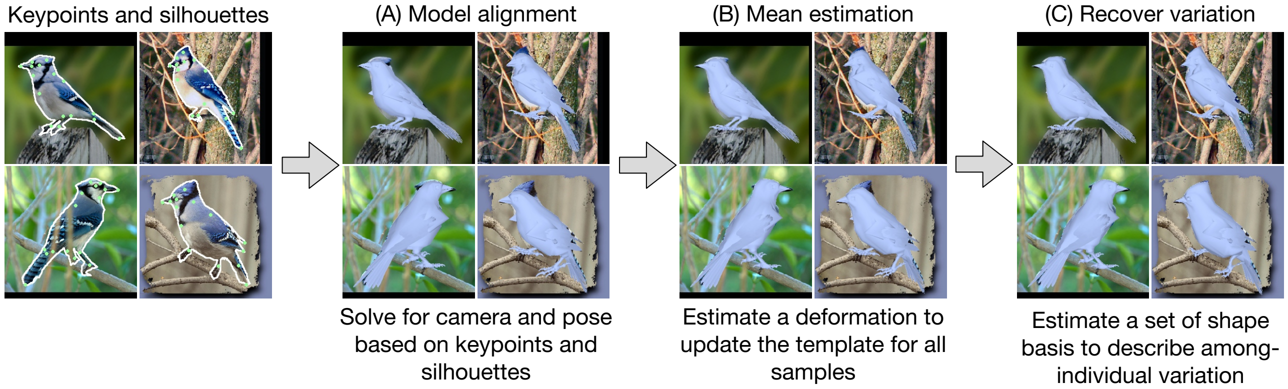
Keypoints and silhouettes
(A) Model alignment
Solve for camera and pose based on keypoints and silhouettes
(B) Mean estimation
Estimate a deformation to update the template for all samples
(C) Recover variation
Estimate a set of shape basis to describe among-individual variation

Figure 2: **Method description**. We align an articulated template to images of a given species (in this case, blue jay), and deform it to first capture the species mean shape (B) and subsequently the shape variations across individuals (C). The mean shape deformations are the same among individuals in the same class, whereas the identity-specific offsets are expressed as linear combinations of a learned blend shape basis.

where ρ is the robust Geman-McClure function [15]. We use the perspective camera model for $\Pi(x)$ and a fixed focal length.

Silhouette reprojection. The silhouette term penalizes the discrepancy between the ground truth mask and the reprojected mask. Here we denote $\mathcal{R}$ as the differentiable rendering of the mesh [43], and $\mathcal{S}$ as the ground truth. Then the silhouette term is

$$E_{msk}(\Theta) = \lambda_{msk} L_\delta(\mathcal{R}(M(\Theta)) - \mathcal{S}). \tag{3}$$

where λ_{msk} is the importance weighting, and L_δ is the smooth $L1$ loss [16].

Prior regularization. We use the means and covariance of α and θ provided by the model, and define the prior term as the squared Mahalanobis distance similar to previous approaches [7, 56, 55] to regularize the optimization.

Figure 2 (A) illustrates the resulting alignments.

3.3. Obtaining species-specific shape

After aligning ABM to each image, we perform the first step of the shape update to better explain the image cues. We define the deformation as a per-vertex displacement vector $\mathbf{dv} \in \mathbb{R}^{3N}$ that updates the template shape as

$$\mathbf{v}_{shape} = \mathbf{v}_{bird} + \mathbf{dv} \tag{4}$$

before different articulations are applied. This deformation vector, shared across all samples, can be seen as a transformation on the original generic shape, to bring it closer to that of the new species.

The species-specific image collection provides guidance with silhouettes and keypoints; $\mathbf{dv}$ will have to explain the discrepancy between the template and some new features presented by the new species. What we hope to achieve is to offset the template to a shape better suited for the new species, which will also condition the next stage when we reconstruct more details for each sample. We call this intermediate shape the species mean shape.

To find the species mean, we fix model parameter Θ and minimize the following objective with respect to $\mathbf{dv}$,

$$E(\mathbf{dv}) = \sum_i E_{kp}^{(i)}(\mathbf{dv}) + E_{msk}^{(i)}(\mathbf{dv}) + E_{sm}(\mathbf{dv}) \tag{5}$$

where E_{kp} and E_{msk} are the keypoint and silhouette reprojection terms from **3.2**, but now influenced by the changing shape $\mathbf{v}_{bird}$ + $\mathbf{dv}$ and are summed over all instances in the collection.

We implement the smoothing term $E_{sm}(\mathbf{dv})$ as

$$E_{sm}(\mathbf{dv}) = E_{edge} + E_{lap} + E_{arap} + E_{sym} \tag{6}$$

E_{edge} smooths displacements between adjacent vertices and is defined as $E_{edge} = \sum_{(p,q)\in edges} \| \mathbf{dv}_p - \mathbf{dv}_q \|_2$. E_{lap} performs Laplacian smoothing on $\mathbf{dv}$ [35]. To preserve local details, E_{arap} enforces as-rigid-as-possible regularization on $\mathbf{v}_{shape}$ [46]. The leg and claw regions are given a higher rigidity. Finally, E_{sym} encourages the overall shape to be symmetrical [55]. A weighted combination of these terms is used to produce the best outcome.

After minimizing the objective with respect to $\mathbf{dv}$, we arrived at an estimated mean, as shown in Figure 2 (B), that explains image evidence better than the generic shape but still needs refinement to fit each individual.

3.4. Reconstructing individuals

Starting from the estimated mean, we aim to reconstruct subject-specific details for each sample. This problem is poorly constrained as we only have one view per subject. However, we can assume that their shapes are drawn from the same distribution. If the estimated mean shape from last step is a good approximation of the species mean, we can

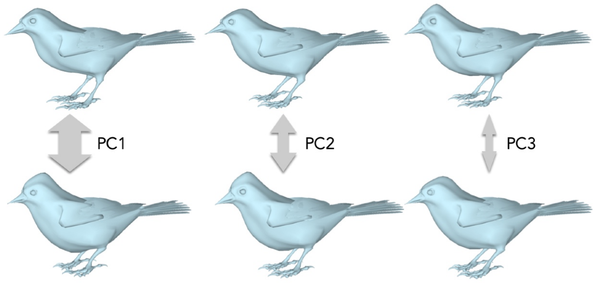

Figure 3: **Principle directions for the blue jay model**, shown with ±1.5std. After learning, we optionally subdivided the surface of the learned models once for higher smoothness; same in Figure 4. Comparison in Sup.Mat.

model the shape variation around the mean with a set of basis vectors.

For each individual i in the collection, its shape can be updated from the last step as

$$\mathbf{v}_{shape}^{(i)} = \mathbf{v}_{bird} + \mathbf{dv} + \sum_{j}^{K} \beta_j^{(i)} * \mathbf{dv}_j \tag{7}$$

where $\{\mathbf{dv}_j\}$ is the set of K basis vectors that describe the variation around the species mean, and $\{\beta_j^{(i)}\}$ are the coefficients for individual i. We can rewrite $\sum_j^K \beta_j^{(i)} * \mathbf{dv}_j$ as $\mathbf{V}\beta^{(i)}$ where $\mathbf{V}$ is the matrix of basis vectors and $\beta^{(i)}$ is the coefficient vector for sample i.

To find $\mathbf{V}$ and $\beta^{(i)}$, we minimize the following objective:

$$E(\mathbf{V}, \beta^{(i)}) = \sum_i [E_{kp}^{(i)} + E_{msk}^{(i)} + E_{sm}^{(i)}](\mathbf{V}, \beta^{(i)}) \tag{8}$$

where E_{kp}, E_{msk} and E_{sm} are defined similarly as in **3.3** but now a function of $\mathbf{V}$ and $\beta^{(i)}$. We also experiment with a soft orthogonality constraint as $\| \mathbf{V}^T\mathbf{V} - I \|_F$ to capture uncorrelated features, and to enforce the magnitude of each basis vector to be constant to resolve the scale ambiguity between the basis and the coefficients. But empirically we do not find improvement in the results.

We obtain a detailed shape for each sample after minimizing the objective. Figure 2 (C) shows results for reconstructing blue jay images (4 of 12 shown).

After reconstruction, we re-learn the mean and the shape basis via PCA to arrive at the final species-specific deformable shape model.

4. Experiments

In this section we empirically show that our method is able to capture realistic shapes and that the recovered shape space is meaningful from a biological standpoint. We also evaluate the accuracy of the methods and the generalization of the learned models. Additionally we incorporate our parametric model in a regression framework that outperforms previous 3D reconstruction methods.

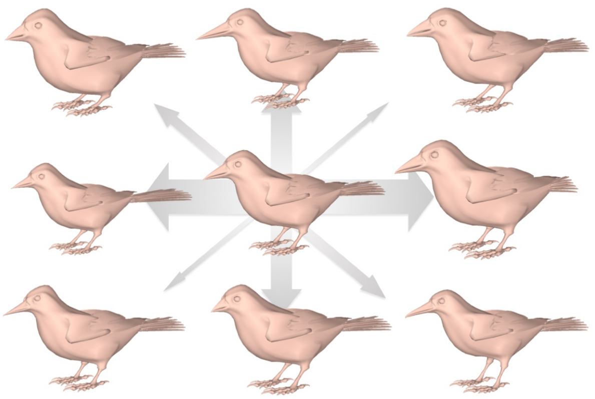

Figure 4: **The first 4 principle directions in the multi-species shape space**. The first two directions are shown with ±1std and the rest with ±2std. Arrow width from thick to thin indicates the first through fourth principle directions, respectively.

We learn the parametric models using the CUB-200 dataset [50]. CUB contains 200 different bird species with segmentation and keypoint annotations. Certain keypoints are not useful for reconstruction, so we adapt only 8 of the original keypoints and additionally hand annotate 10 new keypoints for each sample in our experiments. We select 17 species representing a diverse shape collection. Since the main goal is to accurately capture shapes, we follow previous practice [2, 3, 55] and avoid samples whose pose is difficult to model, including flying and heavily occluded samples. Moreover, if a sample fails the alignment step, we do not include it in the learning steps. On average, we used 18 samples to learn each species. Results in Figure 7 shows realistic captures of different species. More results are included in the supplementary material.

We use blue jays as an example and show variation captured by the species-specific shape basis in Figure 3. We observe variation among individuals in body type, crown size, crown direction and chest.

We also create AVES, a new multi-species avian model by learning a PCA shape space over all species means (we observe similar expressiveness when learning over all individual reconstructions). The learned space, shown in Figure 4, captures characteristics across different avian species. For shape analysis (Fig. 4,5), the PCA space is learned with each sample normalized to have a unit body length so that the analysis is scale invariant. For all other tasks, the PCA is learned without scale normalization.

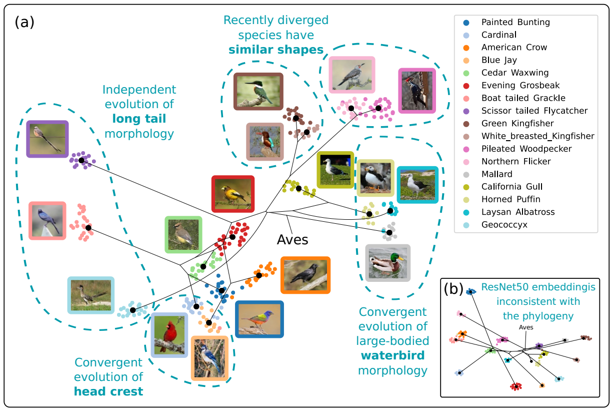

Figure 5: **UMAP** visualization of the principal components of the AVES shape space. (a) Different species are well separated, and similar species are embedded close to each other. Thin lines show the ancestral state reconstruction of bird shape based on the phylogeny. (b) A similar analysis demonstrates that ResNet50 features extracted from the images are inconsistent with the phylogeny.

Recovered shape variation among species is correlated with the phylogeny. We visualize a UMAP [33] embedding of the learned shape PCA in Figure 5. Species are well separated and a phylogenetic analysis [18] shows that recently diverged species are embedded close together. The analysis also reveals several examples of convergent evolution for long tails, waterbird body shape, and head crests. Shape variation across species has high phylogenetic signal and is correlated with the phylogeny, or tree of relatedness [44] (Figure 5, Table 1). On the other hand, visual features extracted using a ResNet50 embedding network trained on CUB [28] are not correlated with the phylogeny, despite clustering well (Table 1). Avian shape captured by our shape space is therefore a more reliable phylogenetic trait than learned perceptual features. Further analyses can be found in the supplementary material.

Features	lambda	p-value
AVES Shape PCs	0.99 ± 0.02	< 0.0001
ResNet50	0.18 ± 0.20	0.60

Table 1: **Captured shapes are correlated with the avian phylogeny.** A trait is consistent with a given phylogeny if it has high phylogenetic signal (Pagel's lambda [39, 40, 44]), which measures the tendency of related species to be closer together in shape space than species drawn at random [9]. Numbers are mean ± std across 100 UMAP embeddings with different initializations for both shape PCs and ResNet50 features.

	Species A		Species B		All 17 species	
	PCK05	IoU	PCK05	IoU	PCK05	IoU
Baseline (ABM)	0.941	0.799	0.809	0.723	0.857	0.765
Species-specific	0.988	0.816	0.963	0.782	**0.954**	**0.805**

Table 2: **Species-specific models** are fitted to the test set of each species. Species A is the painted bunting, which is similar in shape to baseline template. We see a larger performance gain for species that have a very different shape from the template, such as species B, the boat-tailed grackle.

	Species A		Species B		All 17 species	
	PCK05	IoU	PCK05	IoU	PCK05	IoU
Baseline (ABM)	0.940	0.784	0.819	0.723	0.857	0.756
Multi-species	0.969	0.813	0.964	0.793	**0.963**	**0.804**

Table 3: **Multi-species model** learned on $(k-1)$ species and test on the held-out species via k-fold cross-validation. Species A and B are the same as in Table 2.

Comparison of species-specific models with the template model. To evaluate whether the species-specific model can fit to unseen individuals better than the ABM template model (baseline), we split the CUB samples into 70% training and 30% testing for each species. We learn a shape model for each species following the main procedure on the training set. We then fit the model to the testing set by minimizing keypoint and silhouette reprojection errors (optimizing Eq 8 with respect to Θ and β). We evaluate two metrics: the percentage of correct keypoints (PCK) threshold at 5% of the bounding box size, and the intersection over union (IoU) of the silhouettes. Table 2 compares results with the baseline. The species-specific models outperform ABM on all samples for all species, especially for species whose average shape is significantly different from the shape of the template model.

Generalization capability of the multi-species model. To evaluate whether the AVES model can generalize to new species, we conduct k-fold cross-validation where $k = 17$ is the number of species. Every time we learn a new version of AVES on $(k-1)$ species and test it on the held-out species. We fit the model to the held-out species by minimizing the 2D reprojection errors similarly to testing the species-specific models. Table 3 summarizes the results, and shows that the AVES model can indeed generalize to unseen species.

Evaluation on the Aviary dataset. The ABM model was introduced to reconstruct the brown-headed cowbirds in the Penn Aviary dataset [7], which provides multi-view annotations of bird instances across several temporal slices.

We conduct an experiment to see whether the multi-species model can improve over the baseline results. We initialize the shape coefficients to one that matches the de-

Method	PCK05	PCK10	IoU
Cowbird [7]	0.406	0.723	0.605
Ours (AVES)	**0.432**	**0.742**	**0.606**
Cowbird [7] w/ silhouette	0.412	0.731	0.631
Ours (AVES) w/ silhouette	**0.429**	**0.742**	**0.632**

Table 4: **Quantitative evaluation on the Aviary dataset.** We deploy the new shape model on the Aviary dataset to reconstruct examples from multiple views, and compare results to the baseline.

Method	PCK05	PCK10	IoU
CMR [25]	0.432	0.811	0.703
Baseline (ABM) [7]	0.679	0.923	0.706
Ours (AVES)	**0.703**	**0.931**	**0.720**

Table 5: **Quantitative evaluation of different regression methods on the CUB dataset.** We compare the network that predicts the AVES model parameters with the baseline that predicts the ABM parameters and with CMR.

fault shape. We then fixed the coefficient and fit the model to multi-view samples following the baseline procedure described in [7]. To get our results, we allow the shape coefficient to vary towards the end of the optimization. We compare the results in Table 4.

Although the default shape is already a good approximation for cowbirds, we observe improvement in PCKs and comparable results in IoU, indicating that our new shape model AVES better estimates the birds' shapes. We provide additional 3D evaluations of the main approach in the supplementary material.

Regressing 3D bird shape from a single RGB image. To demonstrate the effectiveness of the AVES model, we train a neural network similar to HMR [24] that regresses the AVES model parameters from a single image. The network is supervised using 2D keypoint and silhouette reprojection losses. We compare our approach against a baseline that regresses the ABM [7] parameters and with CMR [24]. Following CMR, we split the test set of CUB into a validation set used for hyperparameter tuning and a separate test set. For more details about the training and the model we refer the reader to the supplementary material. In Figure 8 we show qualitative comparisons. Using a parametric bird model improves the 3D shape over the model-free approach and in turn AVES is able to better capture shape variation across different bird species. Table 5 presents a quantitative comparisons using the available 2D annotations. The AVES model clearly outperforms both the ABM model and CMR on the standard 2D benchmarks.

Application to other animals. Our method is general and can be used to capture other types of animals. We apply it to dogs, which have articulations and intra-breed shape variation, similar to the scenario examined for birds. We use the mean Canis shape from SMAL [56] and the limb scaling implementation from SMBLD [8]. Different than previous works, we do not utilize the shape space. We optimize the pose and limb scales to align the template to images, shown as dark blue in Fig. 6. Given the estimated alignments, we follow our method to capture the shape variation. Fig. 6 shows qualitative results of the Ibizan hounds from Stanford Dogs [8, 27].

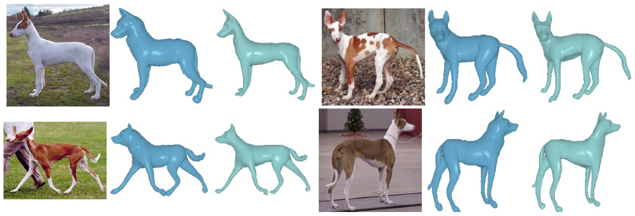

Figure 6: **Application in other animals.** From left to right are images, alignments with SMAL Canis, and our results.

Limitations and failures. We propose a method to capture animal shape space with an articulated template. The quality of the recovered shapes is largely influenced by the accuracy of the estimated pose. Our assumption is that the template can be aligned accurately with the images so that the shape difference is inferred through silhouette difference. Achieving accurate alignment could be difficult for many reasons. First, the pose might not be represented in the pose prior. In our case, the pose of flying birds or lying dogs cannot be solved reliably. Secondly, alignment is ill-defined if the target species has a very different body proportion than the template. The bone parameter of the bird model and the limb scaling of the dog model allow body proportion to adapt and therefore improve alignments. We include some failure cases in the Supplementary Material.

5. Conclusion

In this work, we present a new method for capturing shape models of novel species from image sets. Our method starts from an articulated mesh model and learns intra-species and inter-species deformations using only 2D annotations. Using the captured species-specific shapes, we develop a multi-species statistical shape model, AVES, which is correlated with the avian phylogeny and accurately reconstructs individuals. We use our AVES model for various reconstruction tasks such as model fitting and 3D shape regression from a single RGB image. Our approach focuses on birds but can be applied to other animal classes. Because we use pose priors and predefined kinematic chain, our models cannot capture some extreme pose and shape changes. We leave this challenge for future work.

Acknowledgements: We gratefully appreciate support through the following grants: NSF IIS 1703319, NSF MRI 1626008, NSF TRIPODS 1934960, and NSF CPS 2038873.

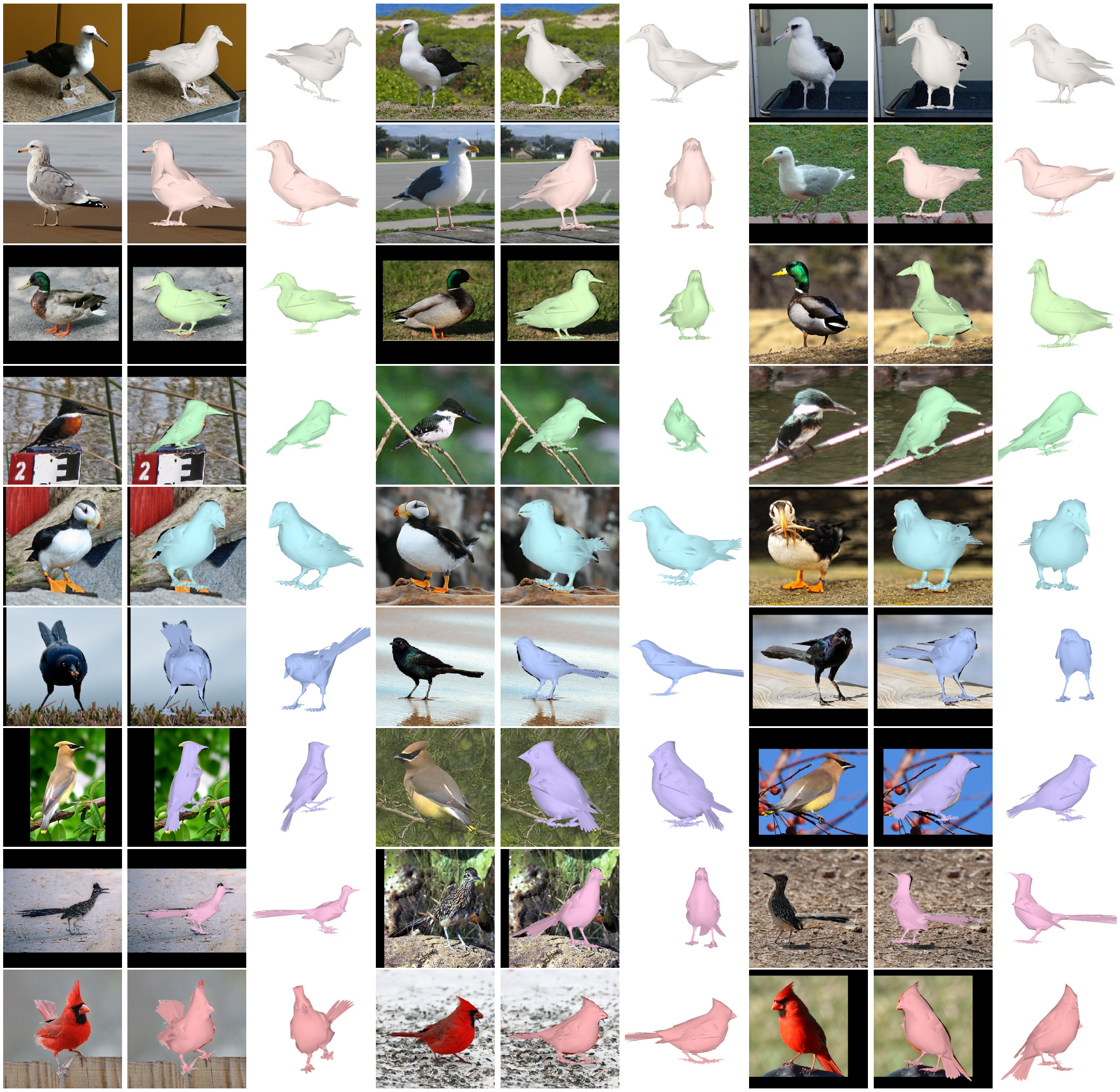

Figure 7: **Examples of learnt species-specific models**. Each row depicts reconstructions using a particular species-specific model that includes identity-specific deformations. Each triplet includes the input image, the reconstructed mesh and the reconstructed mesh from a novel viewpoint.

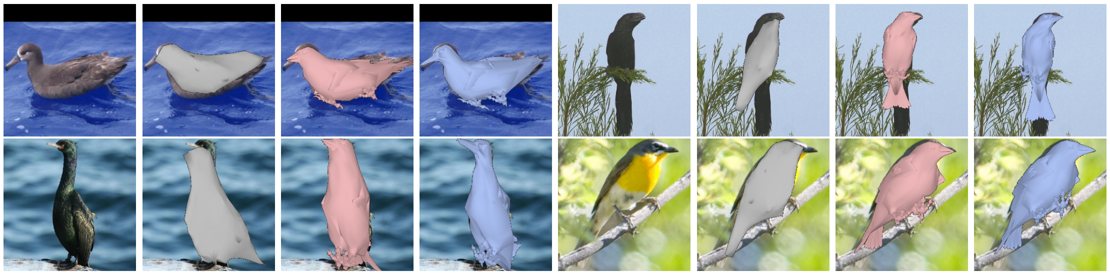

Figure 8: **Qualitative comparison of regression-based methods.** Gray: Reconstruction by CMR [25]. Pink: Baseline (ABM) [7]. Blue: Ours (AVES). More qualitative results can be found in the Sup.Mat.

References

[1] Susan E. Alello and Michael A. Moses. *The Merck Veterinary Manual*. Wiley, 11th edition. 1

[2] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 2, 5

[3] 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. 2, 5

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2

[5] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. *ACM transactions on graphics (TOG)*, 22(3):587–594, 2003. 2

[6] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 2, 3

[7] 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. 1, 2, 3, 4, 6, 7, 8

[8] 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, 2, 3, 7

[9] S. P. Blomberg and T. Garland Jr. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. *Journal of Evolutionary Biology*, 15(6):899–910, 2002. 6

[10] 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. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 2

[11] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 2, pages 690–696. IEEE, 2000. 3

[12] 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. 3

[13] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 3

[14] IL Dryden and KV Mardia. *Statistical analysis of shape*. Wiley, 1998. 3

[15] Stuart Geman. Statistical methods for tomographic image reconstruction. *Bull. Int. Stat. Inst*, 4:5–21, 1987. 4

[16] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4

[17] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. *arXiv preprint arXiv:2007.10982*, 2020. 3

[18] 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. 6

[19] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 3

[20] Tony Heap and David Hogg. Towards 3d hand tracking using a deformable model. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pages 140–145. Ieee, 1996. 3

[21] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *2017 international conference on 3D vision (3DV)*, pages 421–430. IEEE, 2017. 2

[22] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 2

[23] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 2

[24] 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. 2, 7

[25] 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. 3, 7, 8

[26] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1966–1974, 2015. 3

[27] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop*

on *Fine-Grained Visual Categorization (FGVC)*, volume 2. Citeseer, 2011. 7

[28] 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. 6

[29] 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. 2

[30] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. 2

[31] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 3

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 3

[33] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. 6

[34] Desmond Morris. The feather postures of birds and the problem of the origin of social signals. *Behaviour*, 9(2/3):75–113, 1956. 1

[35] Andrew Nealen, Takeo Igarashi, Olga Sorkine, and Marc Alexa. Laplacian mesh optimization. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pages 381–389, 2006. 4

[36] Valsamis Ntouskos, Marta Sanzari, Bruno Cafaro, Federico Nardi, Fabrizio Natola, Fiora Pirri, and Manuel Ruiz. Component-wise modeling of articulated objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2327–2335, 2015. 3

[37] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 2

[38] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. Star: Sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, volume LNCS 12355, pages 598–613, Aug. 2020. 2

[39] Mark Pagel. Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, 26(4):331–348, 1997. 6

[40] Mark Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884, Oct 1999. 6

[41] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2

[42] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018. 2

[43] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 4

[44] Liam J. Revell. Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution*, 1(4):319–329, 2010. 6

[45] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):245, 2017. 3

[46] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116, 2007. 4

[47] Lorenzo Torresani, Aaron Hertzmann, and Chris Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE transactions on pattern analysis and machine intelligence*, 30(5):878–892, 2008. 3

[48] Shubham Tulsiani, Nilesh Kulkarni, and Abhinav Gupta. Implicit mesh reconstruction from unannotated image collections. *arXiv preprint arXiv:2007.08504*, 2020. 3

[49] Sara Vicente and Lourdes Agapito. Balloon shapes: Reconstructing and deforming objects with volume from images. In *2013 International Conference on 3D Vision-3DV 2013*, pages 223–230. IEEE, 2013. 3

[50] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 5

[51] Alexander Weiss, David Hirshberg, and Michael J Black. Home 3d body scans from noisy image and range data. In *2011 International Conference on Computer Vision*, pages 1951–1958. IEEE, 2011. 2

[52] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 2

[53] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, and Kostas Daniilidis. Sparse representation for 3d shape estimation: A convex relaxation approach. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1648–1661, 2016. 3

[54] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5358–5367. IEEE, 2019. 1, 2, 3

[55] Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3955–3963, 2018. 1, 2, 4, 5

[56] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6365–6373, 2017. 1, 2, 3, 4, 7