3D Bird Reconstruction: a Dataset, Model, and Shape Recovery from a Single View

Marc Badger[1,2][0000−0002−6411−706X], Yufu Wang[1,2][0000−0001−9907−8382], Adarsh Modh[1,2][0000−0003−1597−2753], Ammon Perkes[1,2][0000−0001−8932−8309], Nikos Kolotouros[1,2][0000−0003−4885−4876], Bernd G. Pfrommer[1,2][0000−0002−3852−3240], Marc F. Schmidt[1,3][0000−0001−7496−0889], and Kostas Daniilidis[1,2][0000−0003−0498−0758]

[1] University of Pennsylvania, Philadelphia PA 19104, USA
[2] {mbadger, yufu, adarshm, nkolot, pfrommer, kostas}@seas.upenn.edu
[3] {aperkes, marcschm}@sas.upenn.edu

Abstract. Automated capture of animal pose is transforming how we study neuroscience and social behavior. Movements carry important social cues, but current methods are not able to robustly estimate pose and shape of animals, particularly for social animals such as birds, which are often occluded by each other and objects in the environment. To address this problem, we first introduce a model and multi-view optimization approach, which we use to capture the unique shape and pose space displayed by live birds. We then introduce a pipeline and experiments for keypoint, mask, pose, and shape regression that recovers accurate avian postures from single views. Finally, we provide extensive multi-view keypoint and mask annotations collected from a group of 15 social birds housed together in an outdoor aviary. The project website with videos, results, code, mesh model, and the Penn Aviary Dataset can be found at https://marcbadger.github.io/avian-mesh.

Keywords: pose estimation, shape estimation, birds, animals, dataset

1 Introduction

Why computational ethology? Accurate measurement of behavior is vital to disciplines ranging from neuroscience and biomechanics to human health and agriculture. Through automated measurement, computational ethology aims to capture complex variation in posture, orientation, and position of multiple individuals over time as they interact with each other and their environment [1]. Pose trajectories contain rich, unbiased, information from which we can extract more abstract features that are relevant to brain function, social interactions, biomechanics, and health. Studying neural functions in the context of natural social behavior is a critical step toward a deeper understanding how the brain integrates perception, cognition, and learning and memory to produce behavior. Pose trajectories reveal how animals maneuver to negotiate cluttered environments, how animals make decisions while foraging or searching for mates, and how the collective behavior of a group arises from individual decisions. Automated capture

of difficult-to-observe behaviors is transforming diverse applications by streamlining the process of extracting quantitative physiological, behavioral, and social data from images and video.

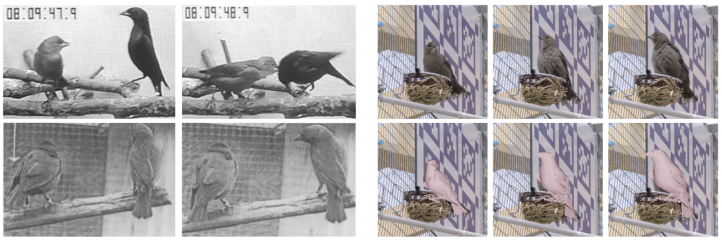

Fig. 1. Visual signals convey important social cues in birds. Motions such as pecking (top left) and wingstrokes (bottom left) drive social behavior in both males and females. Complex wing folding and large changes in body volume when feathers are puffed (upper right) make shape recovery (lower right) a difficult task. Images from [38] and [2].

Why does bird posture matter? Why cowbirds? Understanding how the collective behavior of social groups arises from individual interactions is important for studying the evolution of sociality and neural mechanisms behind social behaviors. Although vocalizations are a clear channel for communication in birds, surprisingly, changes in posture, orientation, and position also play an important role in communication. One of the best studied groups from both behavioal and neuroscience perspectives are the brown-headed cowbirds (*Molothrus ater*). **In cowbirds, females influence the behavior of males through a number of visual mechanisms including "wingstrokes", which involve changes in both pose and shape over time** [38] (Figure 1). Interactions between birds are usually recorded by observing a focal individual's interactions in person in the field. Although insightful, such manual observations contain observer bias, miss interactions between non-focal individuals, and cannot be performed continuously for long periods. Qualitative observations also miss important variation in posture that would be revealed by a quantitative approach. For example, Figure 1 shows changes in pose and shape that can serve as social cues in cowbirds. The ability to estimate the pose of multiple interacting individuals would transform the study of animal communication [1], as is it beginning to do for humans [13, 14, 29]. Estimating the pose and shape of birds in a social context, however, presents several challenges.

Why is estimating bird pose and shape challenging? Recovering shape and pose of birds in the wild is challenging for the following four reasons:

1. Changes in pose and shape are difficult to model in birds.
2. No pose or shape priors are available.
3. Many birds are only visible from a single unoccluded view.
4. Appearance variation in natural settings makes detection difficult.

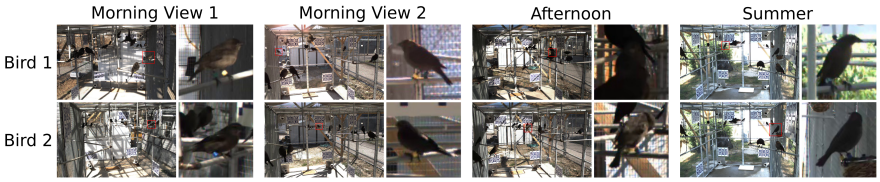

Fig. 2. Appearance variation across bird identity (top vs bottom) and across viewpoint, time of day, and season (1st column vs. columns 2-4 respectively). The red box within the left image of each panel shows the location of the enlarged crop (right image).

Shape is particularly difficult to model because birds have highly mobile feathers that allow dramatic changes in both shape (e.g. tail fanning) and perceived body volume (e.g. feather puffing in Figure 1). Furthermore, when the wings are held next to the body, they are folded in a complex way in which much of the wing surface becomes sandwiched between the top of the wing and the body. These "internal" surfaces cannot be recovered from scans of figurines with folded wings and figurines with wings in intermediate poses are not available. In addition to modeling challenges, cowbirds interact in a complex environment containing extreme variation in illumination and may be heavily occluded either by vegetation by other birds in the group.

Animal posture is often described using joint angles derived from semantic keypoints, joints, or other anatomical locations. This approach is attractive because keypoints are easy to identify and can readily be localized by deep-learning-based software packages such as DeepLabCut [26], DeepPoseKit [10], and LEAP [32]. Under heavy occlusion, however, even multi-view setups frequently do not observe relevant keypoints from more than one view. One solution is to lift the pose from 2D to 3D, but unlike for humans, researchers do not yet have strong species-specific priors for tackling this problem. We overcome the limitations of directly using 2D keypoints and skeletons by fitting a 3D parameterized mesh model with priors learned from a multi-view dataset.

Dataset. With the aim of creating a robust system for estimating the shape and pose of multiple interacting birds over months-long timescales, we recorded the behavior of 15 cowbirds housed together in an outdoor aviary over the course of a single three-month mating season. Our carefully calibrated multi-view dataset contains large variation in (i) bird pose, orientation, and position/depth, (ii) viewpoint across eight cameras, and (iii) appearance across different lighting conditions (time of day and weather) and seasons (Figure 2). Cowbirds have a nearly textureless appearance and birds move freely and explore all three dimensions of their cage, producing a large range of subject depth with respect to the camera. Importantly, both perched and flying birds adopt postures covering a large range of motion in both orientation and pose.

We annotated silhouette and keypoints for 1000 instances and matched these annotations across views. Although 90% of annotated birds were visible from 3

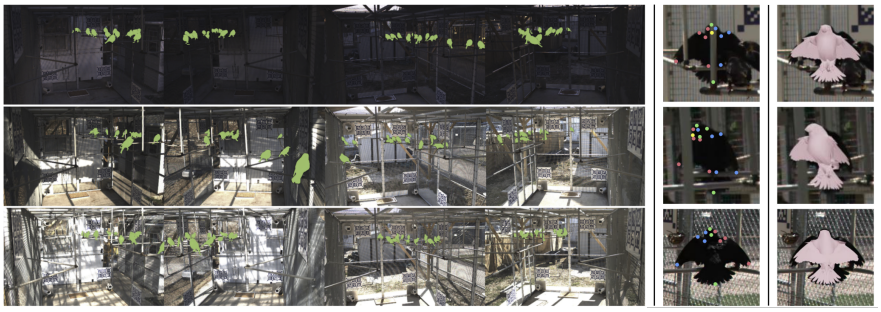

Fig. 3. The dataset and model. We provide multi-view segmentation masks for over 6300 bird instances, keypoints for 1000 bird instances, the first articulated 3D mesh model of a bird, and a full pipeline for recovering the shape and pose of birds from single views.

or more cameras, about half of the annotated instances were occluded to some degree. Only 62% of birds had more than one unoccluded view, highlighting the need for a single-view approach.

After collecting keypoint and silhouette ground truth from multiple views, we fit our avian mesh model using a multi-view optimization-based approach to learn a shape space and pose priors. We then use the model and priors to train a neural network to regress pose parameters directly from keypoint and silhouette data. These pose parameters can be used to initialize a single-view optimization procedure to further refine body pose and shape (Figure 4). We use our dataset for learning instance segmentation and keypoint localization, and for estimating bird pose and shape, but our dataset could also be used in the future for learning Re-ID tasks.

In summary, our contributions are focused around the four challenges mentioned previously:

1. We develop the first parameterized avian mesh model that is capable of capturing the unique pose and shape changes displayed by birds.
2. We fit our mesh model to available multi-view keypoint and silhouette data using an optimization-based approach to obtain an accurate shape space and pose prior.
3. We develop a neural network based pipeline for recovering the shape and pose of birds from a single view.
4. We present a challenging multi-view dataset for studying social behavior in birds. The dataset contains extreme variation in subject appearance and depth and many subjects are fully or partially occluded in all but one view.

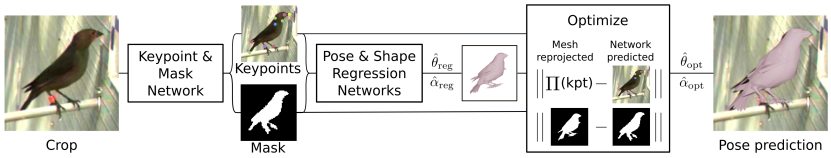

Fig. 4. We estimate the 3D pose and shape of birds from a single view. Given a detection and associated bounding box, we predict body keypoints and a mask. We then predict the parameters of an articulated avian mesh model, which provides a good initial estimate for optional further optimization.

2 Related work

Human pose and shape estimation. Recent advances in human pose estimation have capitalized on i) powerful 2D joint detectors, ii) 3D pose priors, and iii) low-dimensional articulated 3D shape models of the human body. SMPL [25], the most popular formulation, first deforms a template mesh using shape and pose parameters learned from over 1000 registered body scans of people [4] and then uses linear blend skinning (LBS) to transform mesh vertices given a set of joint angles. In SMPLify, Bogo et al. [3] estimate 3D human pose and shape from single images by fitting SMPL to 2D keypoints. Huang et al. [12] extend SMPLify [3] to the multi-view setting and show a positive effect of silhouette supervision in addition to keypoints. Pavlakos et al. [31] estimate pose and shape directly from predicted keypoints and silhouettes in an end-to-end framework. Recent approaches regress pose and shape directly from images and use adversaries with access to a 3D pose dataset [15], Graph-CNN architectures [21], texture consistency [30], and model-fitting within the training loop [20]. All of the above methods base their approach on parameterized mesh models indicating their critical importance for bridging between observation in 2D and estimation in 3D. In contrast to previous works that rely on 3D scans and SMPL-like models to develop meshes and shape spaces for novel domains such as hands [34], faces [22], and four-legged animals [42], we learn our avian mesh model directly from video data of live birds.

Animal pose and shape estimation. Within biology, most work focuses on isolated animals with no background clutter and few occlusions. Mathis et al. [26] and Pereira et al. [32] recently provided tools for training convolutional neural networks for keypoint localization. Graving et al. [10] localize keypoints on three datasets of fruit flies [32], desert locusts [10], and Grévy's zebras [10]. Günel et al. [11] use a Stacked Hourglass network [27] for 2D keypoint localization in flies and perform pictorial structures and belief propagation message passing [7] to reconstruct 3D pose from 2D detections. Liu and Belhumeur et al. [24] use HOG descriptors and linear SVMs to localize bird parts in the more challenging CUB-200-2011 dataset [37]. All of these works are based on the detection and direct triangulation of 2D keypoints. A fundamental challenge, however, is that

any particular keypoint may not be visible from more than one view. Models that constrain the relative position of keypoints, such as the parameterized mesh model we present here, overcome this issue.

Two previous works use articulated graphics models to estimate the pose of flying animals. Fontaine et al. [8] construct a 3D mesh model of a fruit fly and estimate the fly's trajectory and pose over time by fitting the model to three orthogonal views. Breslav [5] create a two-DOF 3D graphics model of a bat and use a Markov Random Field to estimate the 3D pose of bats flying in the wild captured with a multi-view thermal camera setup.

Animal shape estimation is a difficult task. Cashman and Fitzgibbon [6] estimate the shape of dolphins. Ntouskos et al. [28] fit shape primitives to silhouettes of four legged animals. Vincente and Agapito [36] obtain and deform a template mesh using silhouettes from two reference images. Kanazawa et al. [16] learn how animals deform from images by creating an animal-specific model of local stiffness. Kanazawa et al. [17] predict shape, pose, and texture of birds in CUB-200 by deforming a spherical mesh, but do not model pose and thus the locations of wingtips on the mesh are often topologically adjacent to the tail rather than near the shoulders. Zuffi et al. [42] create a realistic, parameterized 3D model (SMAL) from scans of toys by aligning a four-legged template to the scans. They capture shape using PCA coefficients of the aligned meshes and learn a pose prior from a short walking video. Zuffi, Kanazawa, and Black [41] fit the SMAL model to several images of the same animal and then refine the shape to better fit the image data, resulting in capture of both shape and texture (SMALR). Zuffi et al. [40] estimate 3D pose, shape, and texture of zebras in the wild by integrating the SMAL model into an end-to-end network regression pipeline. Their key insight was to first use SMALR to pose an existing horse model and capture a rough texture of the target species. A common feature of these approaches is that they create or leverage a parameterized mesh model. The SMAL model was only trained on four-legged animals so the shape space learned by the model is insufficient for modeling birds, which differ markedly in both limb shape and joint angles. To overcome the lack of a statistical model for birds, we add one additional degree of freedom to each joint and obtain a pose and shape space from multi-view fits to live birds.

Datasets for animal pose estimation. Large-scale object recognition datasets contain many species of animals including dogs, cats, birds, horses, sheep, and more. MS COCO [23] contains 3362 images with bird mask annotations, but no keypoint or pose annotations. The CUB-200 dataset [37] contains 11,788 masks and keypoint instances of birds in the wild. A fruit fly dataset [32] contains 1500 images with centered dorsal views of single flies walking in an arena with a plain white background containing no variation or distractors. The desert locust (800 images) and Grévy's zebras (900 images) include other individuals in the frame, but views are dorsal-only, centered, and narrowly cropped around a focal individual. In contrast our multi-view dataset contains both masks and keypoints of multiple, overlapping subjects and has large variation in relative viewpoint and complex changes in background and lighting.

3 Approach

We adapt a boot-strapped, four-step approach to developing a full pipeline for 3D bird reconstruction from single images (Figure 5). First we develop a parameterized avian mesh and use a multi-view optimization procedure to fit the model to annotations in our dataset. Because they use information from multiple views, these fits are generally good and do not suffer from ambiguities that can plague pose estimation from single views. It is enticing to deploy this multi-view optimization approach towards our end-goal of estimating the pose and shape of all birds over time, but it is slow (initialization is usually far from the target) and requires multiple views in order to produce realistic poses. Nearly 40% of the birds in our dataset were visible from one or fewer unoccluded views, however, indicating the need for a single-view approach. Second, from the multi-view fits, we extract distributions of shape and pose for birds in the aviary, which we use to create a synthetic dataset on which we train neural networks that regress pose and shape parameters from keypoints and silhouettes in a single view. Third, we train a second network to predict an instance segmentation and keypoints given a detection and corresponding bounding box. Finally, we connect the keypoint and segmentation network to the pose regression network. The full pipeline provides a pose and shape estimate from a single view image, which can be used to initialize further optimization (Figure 4).

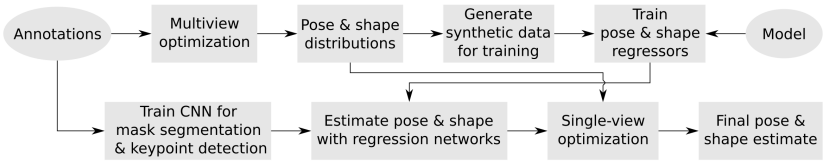

Fig. 5. Overall approach for recovering bird pose and shape from a single view. See Figure 4 for a detailed view of the final pipeline.

Bird detection in full images. We detect bird instances using a Mask R-CNN pretrained on COCO instance segmentation. We removed weights for non-bird classes (leaving bird and background) and then fine-tuned all layers on our dataset for 15 epochs in PyTorch.

Keypoints and silhouette prediction. We train a convolutional neural network to predict keypoints and a silhouette given a detection and corresponding bounding box. We modify the structure of High-Resolution Net (HRNet) [35], which is state-of-the-art for keypoint localization in humans, so that it outputs masks in addition to keypoints. Our modified HRNet achieves 0.46 PCK@05, 0.64 PCK@10, and 0.78 IoU on our dataset.

Skinned linear articulated bird model. To define an initial mesh, joint locations, and skinning weights, we used an animated 3D mesh of a bird model

downloaded from the CGTrader Marketplace website. The model originally contained 18k vertices and 13k faces, but we removed vertices associated with body feathers, eyes, and other fine details to obtain a mesh with 3932 vertices, 5684 faces, and 25 skeletal joints (including a root joint, which is used for camera pose). We use the skinning weights defined in the original file. In addition to skeletal joints, we define 16 mesh keypoints that correspond to the annotated semantic keypoints in our dataset. We obtain keypoint locations by identifying up to four mesh vertices associated with each keypoint and averaging their 3D locations.

To pose the model, we specify a function $M(\alpha, \theta, \gamma, \sigma)$ of bone length parameters $\alpha \in \mathbb{R}^J$ for J joints, pose parameters $\theta \in \mathbb{R}^{3J}$ specifying relative rotation of the joints (and the rotation of the root relative to the global coordinate system) in axis-angle parameterization, global translation inside the aviary γ, and scale σ, that returns a mesh $\mathcal{M} \in \mathbb{R}^{N \times 3}$, with $N = 3932$ vertices. Unlike SMPL [25] and SMAL [42] models, we do not have access to 3D ground truth variation in shape, which prevents the use of shape coefficients drawn from a learned PCA shape space. We mitigate this limitation by including an additional degree of freedom per joint, α_i, that models the distance between parent and child joints, thereby capturing variation in the relative length proportions of the body and limb segments. When birds perch, their wings fold in on themselves and we found that this large deformation is not well modeled by LBS of a single bird mesh model (it is also difficult to capture and register in 3D scans). To overcome this limitation, we use two template poses with identical mesh topology, bones, skinning weights, and keypoints, but with different initial postures: one for birds with their wings outstretched and another for birds with their wings folded (Figure 6). Finally, we also include an overall scale parameter to allow for consistent 3D multi-view estimation among cameras.

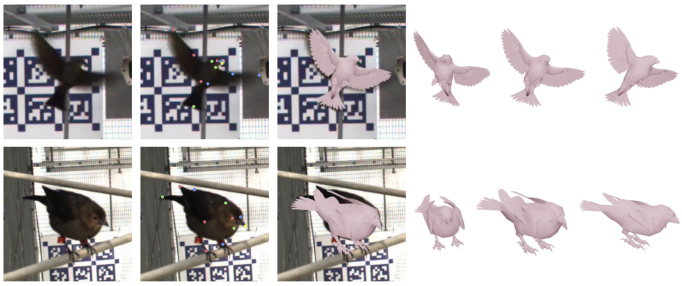

Fig. 6. Our model is capable of capturing both perched and flying poses.

To form the mesh into a given pose, we modify the approach used in SMPL [25] and SMPLify [3] to allow variable bone lengths. Starting with a template mesh $\mathcal{M}^T$ in a canonical pose with joint locations $\mathcal{J} \in \mathbb{R}^{J \times 3}$, we first calculate

the position of each joint i relative to its parent as

$$\mathcal{J}_i^o = \mathcal{J}_i - \mathcal{J}_{\text{parent}(i)}. \tag{1}$$

We then multiply this vector by α_i to adjust the distance between the two joints and form a new skeletal shape $\mathcal{J}'$, still in the canonical pose, with joint locations

$$\mathcal{J}_i' = \alpha_i \, \mathcal{J}_i^o + \sum_{j \in A(i)} \alpha_j \, \mathcal{J}_j^o, \tag{2}$$

where $A(i)$ is the ordered set of joint ancestors of joint i (i.e. all joints encountered moving along the kinematic tree from joint i to the root). Finally, $\mathcal{J}' = J(\alpha)$ is transformed into the final pose using the global rigid transformation $R_\theta(\cdot)$ defined by pose and root orientation parameters θ, and a LBS function $W(\cdot; \mathcal{M}^T)$ is applied. The final mesh vertices are

$$\mathcal{M} = M(\alpha, \theta, \gamma, \sigma) \stackrel{\text{def}}{=} \sigma W(R_\theta(J(\alpha)); \mathcal{M}^T) + \gamma. \tag{3}$$

The positions of 3D keypoints are calculated as $P(M(\alpha, \theta, \gamma))$, where $P(\mathcal{M}) : \mathbb{R}^{N\times 3} \mapsto \mathbb{R}^{K\times 3}$ and K is the number of keypoints. In practice P is simply the average of four selected mesh vertices for each semantic keypoint.

Optimization. To fit our bird model to detected keypoints, we introduce a fitting procedure similar to SMPLify, an optimization-based approach originally described by Bogo et al. [3]. Unlike SMPLify, we capture among individual variation using bone length parameters rather than body shape parameters and we fit to semantic keypoints rather than joint locations. We minimize an objective function with a keypoint reprojection error term and silhouette error term for each camera i, two pose priors, and a prior on the relative 3D distances between joints. Specifically, we minimize:

$$E(\alpha, \theta, \gamma) = \sum_{\text{cam } i} E_{kp}^{(i)}(\cdot;\cdot) + E_{msk}^{(i)}(\cdot;\cdot) + \lambda_\theta E_\theta(\theta) + \lambda_p E_p(\theta) + \lambda_b E_b(\alpha) \tag{4}$$

with

$$E_{kp}^{(i)}(\alpha, \theta, \gamma; K_i, R_i, t_i, \mathcal{P}_i) = \sum_{\text{kpt } k} w_k \rho(\| \, \Pi_{K_i,R_i,t_i}(P(M(\alpha,\theta,\gamma))_k - \mathcal{P}_{i,k} \, \|_2)) \tag{5}$$

and

$$E_{msk}^{(i)}(\alpha, \theta, \gamma; K_i, R_i, t_i, \mathcal{S}_i) = \lambda_{msk} \, \| \, \mathcal{R}_{K_i,R_i,t_i}(M(\alpha,\theta,\gamma)) - \mathcal{S}_i \, \|_2 \, . \tag{6}$$

Equation 5 is a weighted reprojection penalty (using the robust Geman-McClure function ρ [9]) between keypoints $\mathcal{P}_i$ and the projected mesh keypoints $\Pi_{K_i,R_i,t_i}(P(M(\alpha,\theta,\gamma)))$ for pinhole projection function $\Pi(x) = K[R|t]\, x$. The bone lengths, α, are the distances between parent and child joints, θ are the pose parameters, γ is the translation in the global reference frame, K_i, R_i, and t_i are the intrinsics, rotation, and translation, respectively, used in perspective

projection for camera i, and $\mathcal{P}_i$ are the detected or annotated 2D keypoint locations in the image. Equation 6 penalizes differences between an annotated mask $\mathcal{S}_i$ and a rendered silhouette $\mathcal{R}_{K_i,R_i,t_i}(M(\alpha,\theta,\gamma))$ obtained using Neural Mesh Renderer [18]. $E_\theta(\theta) = |\theta - \theta_o|$ is a pose prior that penalizes the L_1 distance from the canonical pose θ_o. $E_p(\theta) = \max(0, \theta - \theta_{\max}) + \max(0, \theta_{\min} - \theta)$ linearly penalizes joint angles outside defined limits $\theta_{\min}$ and $\theta_{\max}$ and $E_b(\alpha) = \max(0, \alpha - \alpha_{\max}) + \max(0, \alpha_{\min} - \alpha)$ penalizes bone lengths outside limits $\alpha_{\min}$ and $\alpha_{\max}$. In the single-view setting, the pose prior (E_θ) and joint angle (E_p) and bone length (E_b) limit losses are disabled and we use the Mahalanobis distance to the distribution of multi-view pose and shape estimates instead. We minimize the objective in 4 using Adam [19] in PyTorch.

Synthetic data and pose and shape regression. After performing multi-view optimization on 140 3D bird instances in our annotated dataset, we fit a multivariate Gaussian to the estimated pose parameters (pose, viewpoint, and translation). We then sample 100 random points from this distribution for each bird instance, project the corresponding model's visible keypoints onto the camera and render the silhouette, generating 14,000 synthetic instances for training. We keep the bone lengths of the original 140 instances, but add in random noise to the bone lengths for each sample.

We train pose and shape regression networks on the 14,000 synthetic single-view instances supervised by the ground truth pose and shape parameters. For the pose regression network inputs are 2D joint locations and targets are 3D rotations, which are first transformed to the representation proposed by Zhou et al. [39] before computing the L^2 loss. The pose regression network is an MLP with two fully connected layers with the final layer outputting $25 * 6 + 3$ translation parameters. The shape regression network takes in a mask and contains one 5×5 convolutional layer followed by four 3×3 convolutional layers and a fully connected layer with 24 outputs, corresponding to the 24 bone lengths. Each convolutional layer is followed by batch normalization and max-pooling layers. Training was performed for 20 epochs using Adam.

4 The cowbird dataset

Image acquisition and aviary details. We captured video of 15 individual cowbirds (*Molothrus ater*) in an outdoor aviary from March to June using eight synchronized cameras recording 1920×1200 images at 40 Hz. The aviary is 2.5 meters in height and width and is 6 meters long. Cameras were positioned in the corners and oriented so that their combined fields view provided maximal coverage of the aviary volume by at least four cameras. Intrinsic parameters were estimated for each camera using a standard checkerboard and the camera calibration package in ROS. Extrinsic parameters for camera orientation and translation were estimated online via the TagSLAM package [33] using arrays of fiducial markers permanently attached to the aviary walls.

Dataset annotation and statistics. From the above recordings, we exported sets of synchronous frames from 125 "moments" (1000 images) drawn

from 10 days uniformly distributed over the recording period (an average of 12.5 uniformly distributed moments each day). On all images, we exhaustively annotated instance segmentation masks for all visible birds, producing over 6355 masks and bounding boxes. On a subset of 18 moments across six of the 10 days we also annotated the locations of 12 semantic keypoints on a total of 1031 masks (Figure 3). We annotated the bill tip, right and left eyes, neck, nape, right and left wrists, right and left wing tips, right and left feet, and the tail tip. Statistics on the visibility of keypoints (Table S7) and a comparison with other animal datasets (Tables S4, S5) are in the supplementary material.

We manually associated keypoint annotations within each moment across camera views to create 3D instance ID tags. From the 3D instance ID tags, 64%, 26%, and 10% of birds were fully or partially visible from four or more cameras, three cameras, and two or fewer cameras, respectively (Supplementary Table S6). The average width $\times$ height of bird masks was 68×75 pixels (or $\approx$ 5% of image width; the 5th and 95th percentiles of bird max dimensions were 17×19 and 239×271 pixels, respectively). We provide four types of test/train splits: by moment, by day, by time of day (morning vs. afternoon), and by season (March and April vs May and June). Birds wore colored bands on their legs that, when visible, could provide the true ID of the bird, but we leave the potential application of this dataset to the Re-ID task for future work.

5 Experiments

Detection. We first evaluate the performance of Mask R-CNN on instance segmentation of birds using our dataset. We show excellent generalization (AP = 0.52) when predicting masks on unseen days in the test set (Figure 7). Further analyses and performance on additional splits of the dataset (e.g. split by time of day or season) are provided in Supplementary Table S1.

Fig. 7. Instance detections made by a fine tuned Mask R-CNN network over a large range of lighting conditions and views. Best viewed in color.

Multi-view optimization. We fit our articulated avian mesh model to annotations corresponding to each 3D bird instance in our keypoint dataset. We fit using all keypoint labels from all available views. We present qualitative results in Figure 8. Our fitting procedure resulted in many plausible results but also in many failure cases, shown in the bottom row of Figure 8. From the multi-view

fits, we obtained a pose and shape space for the mesh model, which we display in the supplementary video. We perform an ablation experiment to investi-

Fig. 8. Multi-view optimization-based fits of the bird mesh to keypoint and mask annotations in our dataset (upper section). Failure cases are shown in the lower section.

gate the effects of pose priors and joint and bone limits on performance in the single-view setting. For each ablation, we remove the corresponding term from the objective and report its effect on the accuracy of projected mesh keypoints and silhouettes, which we report in Supplementary Table S3. We measure keypoint accuracy relative to ground truth keypoints using PCK at two thresholds calculated based on the largest dimension of the bounding box and we measure the accuracy of the projected silhouettes using IoU with ground truth masks. We budget 500 iterations for fitting each instance for all settings. The PCK increased as we removed the pose prior and bone limit (but not pose limit) terms from our objective. This increase indicates the model is achieving a better fit to the keypoints, potentially at the cost of producing an unrealistic fit, as might be indicated by the simultaneous decrease in IoU as priors are removed.

Do silhouettes improve multi-view optimization? We compared fits of the model with and without the silhouette term (Equation 6) in the objective. The silhouette term improves IoU while only slightly affecting keypoint error (Table 1). More importantly, the silhouette term allows the model to better capture changes in shape produced during feather puffing (Figure 1).

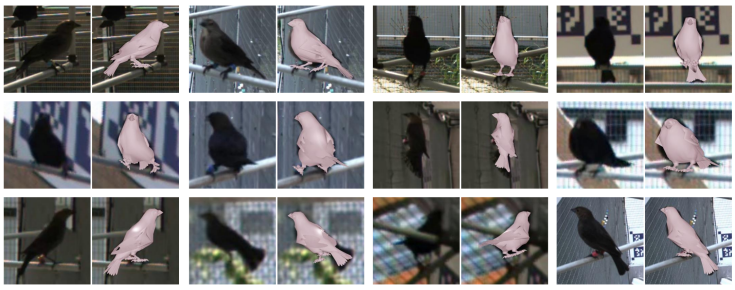

Fig. 9. Regression-based recovery of bird pose and shape from a single view. Each panel shows the input image and refined mesh (see Figure 4).

Table 1. Ablation study of the silhouette term in the multi-view optimization setting. PCK@05 and PCK@10 denote percent correct keypoints within 5% and 10% of bounding box width, respectively. Silhouettes improve IoU with minimal effect on keypoint error.

	weight ratio (kpt:mask)	PCK@05	PCK@10	IoU
keypoints only	N/A	0.356	0.631	0.540
keypoints + mask	10:1	0.355	0.637	0.560
keypoints + mask	1:1	0.328	0.618	0.624

3D shape and pose recovery from a single view. Our single-view pipeline produces poses that are consistent across views (Table 2, Supplementary Figure S1). To overcome scale ambiguity, we fix pose and shape and then find the Procrustes transformation (translation, rotation, and scaling) that minimizes keypoint reprojection error in each additional view. We also perform experiments to evaluate the individual components of our full pipeline (Table 3). We first compare pose regression alone (i.e. not optimizing after regression), single-view optimization alone (i.e. not initialized by pose regression network), and the full pipeline. Although the regression network alone is less "accurate" than single-view optimization (Table 3), the pose regression network produces good estimates of global pose, which allows optimization to proceed much faster. Additional examples are shown in Figure 9. Finally, we demonstrate that our model and bone length formulation generalize to similar bird species in the CUB-200 dataset (Supplementary Figure S3).

Failure cases. Occasional failures resulted in unnatural poses, which are shown in Supplementary Figure S2. To evaluate the cause of these failures, two annotators inspected the same random sample of 500 crops and rated their confidence in each bird's pose (confident, semi-confident, not-confident). They then rated the predicted keypoints as good or bad for all crops. Finally, they viewed the mesh fits and rated each as a success or failure. We found that 84% of confident, 35% of semi-confident, and 12% of not-confident crops were fit

Table 2. Cross-view PCK and IoU of projected meshes from the single-view pipeline. Values are averaged across all views except the view used to obtain the mesh. Ground truth pipeline input means the keypoint and mask network predictions (Figure 4) are replaced by ground truth annotations.

Pipeline input	PCK@05	PCK@10	IoU
predictions	0.313	0.632	0.589
ground truth	0.332	0.635	0.586

Table 3. Same-view evalutation of the single-view pipeline and ablations. Regression and optimization are performed using keypoint and mask predictions and evaluated against ground truth. Additional results are presented in Supplementary Table S2.

	PCK@05	PCK@10	IoU
regression	0.104	0.318	0.483
optimization	0.331	0.575	0.641
reg. + opt.	0.364	0.619	0.671

successfully. Bad keypoint detection was responsible for 60% of failures. Even good fits are not perfect, particularly in the tail and feet. Adding more degrees of freedom to the model, such as tail fanning, and annotating additional keypoints on the toes would improve these areas.

6 Conclusions

We present an articulated 3D model that captures changes in pose and shape that have been difficult to model in birds. We provide a novel multi-view dataset with both instance masks and keypoints that contains challenging occlusions and variation in viewpoint and lighting. Our single-view pipeline recovers cross-view-consistent avian pose and shape, and enables robust pose estimation of birds interacting in a social context. We aim to deploy our pipeline in the aviary to better understand how individual interactions drive the formation of avian social networks.

An interesting feature of birds is that variation in a single individual's shape across time can be much larger than overall shape variation among individuals (e.g. due to feather fluffing shown in Figure 1). In the future, it will be interesting to apply our pipeline to video data and additional species to develop a more nuanced model of how shape varies across time, individuals, and species.

Capturing 3D pose is critical to understanding human and animal health and behavior. Pose data produced by our pipeline will be useful for addressing how flying animals maneuver, negotiate cluttered environments, and make decisions while foraging or searching for mates, and how the collective behavior of a group arises from individual decisions.

Acknowledgements. We thank the diligent annotators in the Schmidt Lab, Kenneth Chaney for compute resources, and Stephen Phillips for helpful discussions. We gratefully acknowledge support through the following grants: NSF-IOS-1557499, NSF-IIS-1703319, NSF MRI 1626008, NSF TRIPODS 1934960.

References

1. Anderson, D.J., Perona, P.: Toward a science of computational ethology. Neuron **84**(1), 18 – 31 (2014). https://doi.org/https://doi.org/10.1016/j.neuron.2014.09.005, http://www.sciencedirect.com/science/article/pii/S0896627314007934
2. Baillie, K.U., Spitzer, S., Crucius, D.: 'Smart aviary' poised to break new ground in behavioral research (2019), https://penntoday.upenn.edu/news/smart-aviary-poised-break-new-ground-behavioral-research
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Computer Vision – ECCV 2016. pp. 561–578. Lecture Notes in Computer Science, Springer International Publishing (Oct 2016)
4. Bogo, F., Romero, J., Loper, M., Black, M.J.: FAUST: Dataset and evaluation for 3D mesh registration. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, Piscataway, NJ, USA (Jun 2014)
5. Breslav, M.: 3D pose estimation of flying animals in multi-view video datasets. Ph.D. thesis, Boston University (2016)
6. Cashman, T., Fitzgibbon, A.: What shape are dolphins? Building 3D morphable models from 2D images. IEEE Transactions on Pattern Analysis and Maching Intelligence **35**, 232 (January 2013), https://www.microsoft.com/en-us/research/publication/shape-dolphins-building-3d-morphable-models-2d-images/
7. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. Int. J. Comput. Vision **61**(1), 5579 (Jan 2005). https://doi.org/10.1023/B:VISI.0000042934.15159.49, https://doi.org/10.1023/B:VISI.0000042934.15159.49
8. Fontaine, E.I., Zabala, F., Dickinson, M.H., Burdick, J.W.: Wing and body motion during flight initiation in drosophila revealed by automated visual tracking. Journal of Experimental Biology **212**(9), 1307–1323 (2009). https://doi.org/10.1242/jeb.025379, https://jeb.biologists.org/content/212/9/1307
9. Geman, S., McClure, D.: Statistical methods for tomographic image reconstruction. Bulletin of the International Statistical Institute **LII**(4), 5–21 (1987)
10. Graving, J.M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B.R., Couzin, I.D.: DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. eLife **8**, e47994 (2019)
11. Günel, S., Rhodin, H., Morales, D., Campagnolo, J., Ramdya, P., Fua, P.: DeepFly3D, a deep learning-based approach for 3D limb and appendage tracking in tethered, adult *Drosophila*. eLife **8**, e48571 (2019)
12. Huang, Y., Bogo, F., Lassner, C., Kanazawa, A., Gehler, P.V., Romero, J., Akhter, I., Black, M.J.: Towards accurate marker-less human shape and pose estimation over time. In: 2017 International Conference on 3D Vision (3DV). pp. 421–430 (2017)
13. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 3334–3342 (2015)
14. Joo, H., Simon, T., Cikara, M., Sheikh, Y.: Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10865–10875 (2019)

15. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Recognition (CVPR) (2018)
16. Kanazawa, A., Kovalsky, S., Basri, R., Jacobs, D.: Learning 3D deformation of animals from 2D images. Computer Graphics Forum **35**(2), 365–374 (2016). https://doi.org/10.1111/cgf.12838, https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12838
17. Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J.: Learning category-specific mesh reconstruction from image collections. In: ECCV (2018)
18. Kato, H., Ushiku, Y., Harada, T.: Neural 3D mesh renderer. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3907–3916 (2018)
19. Kingma, D.P., Ba, J.L.: Adam : A method for stochastic optimization. arXiv (2014)
20. Kolotouros, N., Pavlakos, G., Black, M., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2252–2261 (2019)
21. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4496–4505 (2019)
22. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Trans. Graph. **36**(6) (Nov 2017). https://doi.org/10.1145/3130800.3130813, https://doi.org/10.1145/3130800.3130813
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014)
24. Liu, J., Belhumeur, P.N.: Bird part localization using exemplar-based models with enforced pose and subcategory consistency. In: 2013 IEEE International Conference on Computer Vision. pp. 2520–2527 (2013)
25. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
26. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M.: DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience **21**(9), 1281–1289 (2018)
27. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 483–499. Springer International Publishing, Cham (2016)
28. Ntouskos, V., Sanzari, M., Cafaro, B., Nardi, F., Natola, F., Pirri, F., Ruiz, M.: Component-wise modeling of articulated objects. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 2327–2335 (2015)
29. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10967–10977 (2019)
30. Pavlakos, G., Kolotouros, N., Daniilidis, K.: Texturepose: Supervising human mesh estimation with texture consistency. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 803–812 (2019)
31. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3D human pose and shape from a single color image. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 459–468 (2018)

32. Pereira, T.D., Aldarondo, D.E., Willmore, L., Kislin, M., Wang, S.S.H., Murthy, M., Shaevitz, J.W.: Fast animal pose estimation using deep neural networks. Nature Methods **16**, 117–125 (2019)
33. Pfrommer, B., Daniilidis, K.: Tagslam: Robust slam with fiducial markers. arXiv (2019)
34. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Trans. Graph. **36**(6) (Nov 2017). https://doi.org/10.1145/3130800.3130883, https://doi.org/10.1145/3130800.3130883
35. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5686–5696 (2019)
36. Vicente, S., Agapito, L.: Balloon shapes: Reconstructing and deforming objects with volume from images. In: 2013 International Conference on 3D Vision - 3DV 2013. pp. 223–230 (2013)
37. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
38. West, M.J., King, A.P.: Female visual displays affect the development of male song in the cowbird. Nature **334**, 224–246 (1988)
39. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5738–5746 (2019)
40. Zuffi, S., Kanazawa, A., Berger-Wolf, T., Black, M.: 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). pp. 5358–5367 (2019)
41. Zuffi, S., Kanazawa, A., Black, M.J.: Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3955–3963 (2018)
42. Zuffi, S., Kanazawa, A., Jacobs, D.W., Black, M.J.: 3D menagerie: Modeling the 3D shape and pose of animals. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5524–5532 (2017)