Supplementary Material for "3D Bird Reconstruction: a Dataset, Model, and Shape Recovery from a Single View"

Marc Badger[1,2], Yufu Wang[1,2], Adarsh Modh[1,2], Ammon Perkes[1,2], Nikos Kolotouros[1,2], Bernd G. Pfrommer[1,2], Marc F. Schmidt[1,3], and Kostas Daniilidis[1,2]

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

Introduction

In the supplementary material, we present a supplementary video and the following additional material:

1. Cross-view examples and failure cases from the single-view pipeline
2. Application of our pipeline to other bird species
3. An assessment of the performance of Mask R-CNN on various dataset splits
4. Ablation experiments and additional evaluations
5. A comparison of our dataset with other animal datasets
6. Statistics for mask and keypoint annotations in our dataset.

1 Cross-view examples and failure cases produced by the single-view pipeline

Our single-view pipeline produces poses that are consistent across views (Table 2 in the main paper). In Figure S1 we present visual examples of meshes projected onto views that were not used to obtain the mesh. The distributions of pose and shape obtained from multi-view optimization provide a sufficient prior for single-view pose estimation.

The keypoint detector sometimes fails completely for difficult poses (Figure S2, top row) or swaps left and right keypoints (bottom row), which cause a faulty prediction by the pose regression network. Sixty percent of failure cases were associated with bad keypoint detection. Even in cases with very few keypoints, however, the pipeline produces meshes that are consistent with the silhouette.

2 Application to other bird species in CUB-200

We also predict pose and shape for several other bird species (Figure S3). We annotated our set of keypoints (see main text) on several examples from CUB-200

Fig. S1. Three examples of single-view fits (source views) visualized from additional views that were not used to obtain the mesh (cross-view examples).

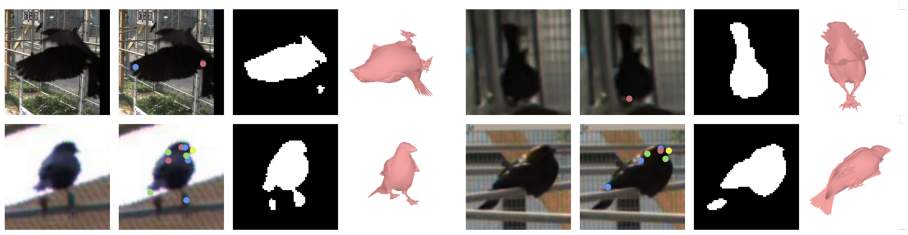

Fig. S2. Failure cases from the full pipeline. From left to right, each panel shows the input image, predicted keypoints, predicted mask, regressed mesh, and refined mesh.

and input the mask and keypoints into our single-view pipeline, starting at the pose and shape regression networks. The pose regression networks and bone length formulation of shape variation allow successful fits on several similar species.

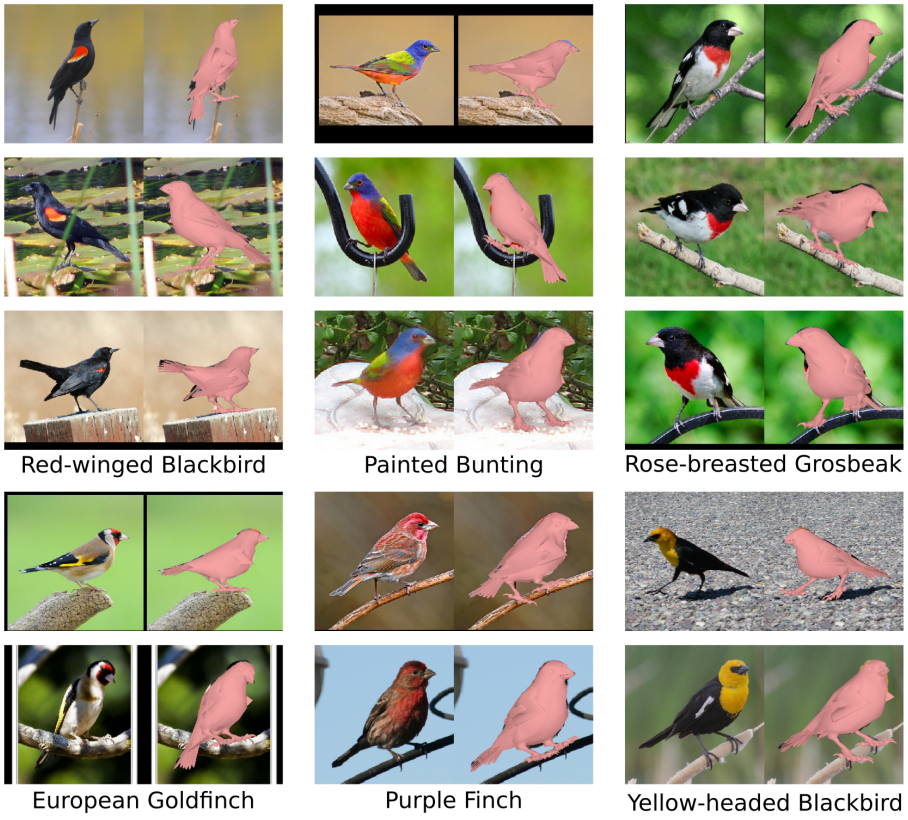

Fig. S3. Our mesh, pose regression networks, and single-view optimization procedure generalize to similar bird species in CUB-200 using distributions of shape and pose extracted from our multi-view dataset.

3 Performance of Mask R-CNN

Here we evaluate the performance of Mask R-CNN on instance segmentation of birds on additional splits of the dataset. We train and test on datasets split 1) randomly by moment, where each set of images from the eight camera views is randomly assigned to train/test, 2) randomly by day, where all samples from a given recording day are—together—assigned randomly to train/test, 3) by

the time of day, where we train/test on data before noon and test/train on data after noon, and 4) by season, where we train/test on data from the spring (March and April) and test/train on data from the summer (May and June) (Table S1). Performance was generally highest when data was split by moment, the most granular split.

Table S1. Average precision of Mask R-CNN predictions from 50% to 95% IoU (AP), precision at 50% and 75% IoU (AP50 and AP75, respectively), and average precision across small ($0 - 32^2$ pixels), medium ($32^2 - 96^2$ pixels), and large (over 96^2 pixels) annotations (APs, APm, and APl, respectively) in the multi-view setting. Given the size categories above, 33% of annotations were small, 59% were medium, and 8% were large. The training set is indicated by "tr." if applicable.

Split	Num. train	Num. test	AP	AP50	AP75	APs	APm	APl
by moment	695	215	0.59	0.89	0.70	0.38	0.64	0.80
by day	595	315	0.53	0.82	0.62	0.35	0.58	0.71
by time of day (tr. AM)	486	424	0.58	0.90	0.68	0.38	0.64	0.80
by time of day (tr. PM)	424	486	0.54	0.86	0.63	0.41	0.59	0.72
by season (tr. spring)	427	483	0.58	0.91	0.69	0.42	0.63	0.79
by season (tr. summer)	483	427	0.52	0.81	0.60	0.35	0.58	0.73

4 Ablation experiments and additional evaluations

Single-view pipeline. Here we extend our evaluation of our single-view pipeline (Table 3 in the main paper) by replacing keypoint and mask network predictions with ground truth annotations. Using ground truth annotations, pose regression followed by optimization produces high quality fits with PCK@10 ¿ 0.96 (Table S2). Thus more accurate keypoint detection would significantly improve the performance of our full pipeline. However, cross-view evaluations given in Table 2 in the main paper demonstrate that these improvements may not translate to better overall reconstruction.

Optimization-based multi-view pose estimation. We also perform ablation experiments to investigate the effects of pose priors and joint and bone limits on performance in the multi-view setting (Table S3). We find that PCK increases as the pose prior and bone limits (but not pose limits) are removed. IoU decreases significantly, indicating that these improvements may be the result of unrealistic fits.

5 Comparison with other animal datasets

With the exception of Wah et al. [5] (CUB-200), datasets of animal shape and pose are extremely limited. Although other datasets contain more images, more

Table S2. Same-view evaluation of the single-view pipeline and ablations. In the upper section, regression and optimization are performed using keypoint and mask predictions and evaluated against ground truth. In the lower section, regression and optimization are performed using keypoint and mask ground truth annotations and evaluated against the same. Data are mean ± standard error. The upper section is from the main paper and is reproduced here for comparison.

	PCK@05	PCK@10	IoU
regression using predictions	0.104 ± 0.014	0.318 ± 0.027	0.483 ± 0.011
optimization using predictions	0.331 ± 0.025	0.575 ± 0.030	0.641 ± 0.014
reg. + opt. using predictions	0.364 ± 0.028	0.619 ± 0.031	0.671 ± 0.014
regression using ground truth	0.135 ± 0.016	0.357 ± 0.031	0.476 ± 0.013
optimization using ground truth	0.783 ± 0.020	0.933 ± 0.011	0.646 ± 0.012
reg. + opt. using ground truth	0.825 ± 0.021	0.967 ± 0.008	0.696 ± 0.011

Table S3. Comparison of PCK and IoU of projected mesh with ground truth keypoint and mask annotations for the multi-view setting. PCK@05 and PCK@10 denote percent correct keypoints within 5% and 10% of bounding box width, respectively. Results presented here are fit without the silhouette term in the objective.

	PCK@05	PCK@10	IoU
Full	0.357	0.623	0.541
$-E_\theta$ (pose prior)	0.417	0.677	0.532
$-E_p$ (pose limits)	0.361	0.643	0.538
$-E_b$ (bone limits)	0.383	0.657	0.466
E_{kp} only	0.511	0.724	0.413

masks, or more keypoints, no other dataset contains both masks and keypoints of multiple interacting subjects against a complex background (Tables S4 and S5).

Table S4. Existing datasets for animal pose estimation. Our multi-view animal pose dataset contains both masks and keypoints, has large variation in both relative viewpoint (including subject depth) and lighting, and has multiple instances per image and a complex background. Wah et al. [5] is the CUB-200 dataset.

	Animal	Max res.	Images	Viewpoint(s)
Wah et al. [5]	Bird	500×500	11,788	single (varying)
Breslav et al. [1]	Moth	600×800	800	single (behind)
Pereira et al. [4]	Fly	192×192	1500	single (above)
Graving et al. [2]	Locust	160×160	800	single (above)
Graving et al. [2]	Zebra	160×160	900	single (above)
Günel et al. [3]	Fly	512×256	11,063	7 cams (ground plane)
Ours	Cowbird	1920×1200	1000	8 cams (3D volume)

Table S5. Existing datasets for animal pose estimation (continued). "Inst. × keypoints" is the number of instances with keypoint labels × the number of keypoints labeled per instance. "Mult. Inst." is whether there are multiple subjects in any given image.

	Animal	Masks	Inst. × keypoints	Mult. Inst.	Background
Wah et al. [5]	Bird	11,788	$11,788 \times 15$	no	complex
Breslav et al. [1]	Moth	800	400×4	no	plain
Pereira et al. [4]	Fly	0	1500×32	no	plain
Graving et al. [2]	Locust	0	800×35	yes	plain
Graving et al. [2]	Zebra	0	900×9	yes	complex
Günel et al. [3]	Fly	0	$11,063 \times 19$	no	plain
Ours	Cowbird	6355	1031×12	yes	complex

6 Multi-view cowbird and keypoint visibility

Over the 16 moments with keypoint annotations, we recovered 237 three-dimensional bird instances. The visibility of these instances by the cameras in the aviary is shown in Table S6. Mean reprojection error for all keypoints was 9.0 pixels and ranged between 8.1 pixels for the bill tip to 9.5 pixels for the wing tips. The tail tip was seen by the highest number of cameras on average across 3D instances (3.4 views per bird) and the eyes were seen by the fewest cameras (2.0 views per bird).

Table S6. Visibility of 3D bird instances by cameras. The entry in each column is the number (percent) of 3D bird instances that are visible from the corresponding number of cameras. An annotation was labeled as unoccluded if all of the bird was visible and no parts were hidden behind other birds or structures in the environment.

Visible from:	0 cams	1 cam	2 cams	3 cams	4 cams	5 cams
All annotations	0 (0.00)	7 (0.03)	16 (0.07)	62 (0.26)	123 (0.52)	29 (0.12)
Unoccluded annotations	29 (0.12)	61 (0.26)	78 (0.33)	42 (0.18)	27 (0.11)	0 (0.00)

Table S7. Reprojection error of keypoint annotations and visibility of anatomical landmarks. "Num. cams" indicates the mean number of cameras that viewed each keypoint across all 3D bird instances. On average, birds were viewed by 3.6 cameras (and had 3.6 corresponding two-dimensional keypoint annotations).

Keypoint	Reprojection error (pixels)	Num. cams
Bill tip	8.1	2.9
Left eye	9.0	2.1
Right eye	8.7	2.0
Neck	8.4	2.6
Nape	9.5	3.0
Left wrist	9.3	2.5
Right wrist	9.4	2.5
Left wing tip	9.4	2.7
Right wing tip	9.7	2.7
Left foot	9.1	2.7
Right foot	8.8	2.6
Tail tip	8.5	3.4

References

1. Breslav, M.: 3D pose estimation of flying animals in multi-view video datasets. Ph.D. thesis, Boston University (2016)
2. 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)
3. 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)
4. 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)
5. 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)