# Quantifying Bird Skeletons

Zhizhuo Zhou[1], Gemmechu Hassena[2], Brian C. Weeks[3], David F. Fouhey[1]
[1] University of Michigan EECS, [2] Addis Ababa University, [3] University of Michigan SEAS

## Abstract

*Birds have well-known distributions, phylogenetic relationships, and life histories, making them a powerful model system for understanding biotic responses to global environmental change. However, there are hundreds of thousands of museum skeletal specimens that could be analyzed to further our understanding of avian responses to climate change, but remain under-utilized due to the practical constraints of measuring elements of skeletons by hand. We introduce a dataset and system for measuring skeletal traits from museum specimens that reduces capture time by 15x for an initial effort to measure 10 traits per skeleton, allows for post hoc addition of trait data orders of magnitude faster than traditional methods, and shows high accuracy even when trained with limited data.*

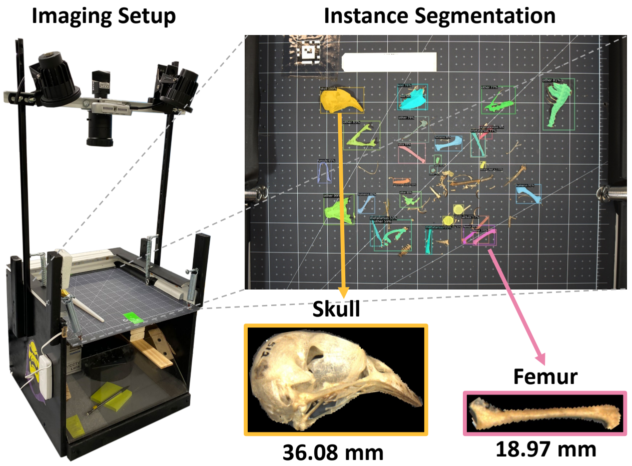

Figure 1. Overview of the system: the imaging setup, an instance segmentation output, and processed bone measurements.

## 1. Introduction and Related Work

Climate change is predicted to result in changes in animal body size, with warming temperatures expected to result in size reductions broadly across the tree of life [7]. Size impacts nearly every aspect of an organism's ecology, thus understanding how species compensate for warming-driven size reductions is key to understanding and predicting the impacts of climate change on the world's ecosystems.

In birds, recent warming has driven consistent reductions in body size, with contemporaneous changes in body shape [19]. While it is generally known that morphology is predictive of ecology [16], key to understanding the impacts of climate change on animal ecology and fitness is the collection of extensive and detailed functional trait data from a large number of individuals within species through time. Museum skeletal specimens could fill this data gap: over 500,000 bird skeletal specimens are held in natural history museums. However, measuring only a limited number of traits from a specimen can take a well-trained technician 15-30 minutes; as a result, collecting only a small number of species by hand at the timescales necessary to understand biotic responses to climate change can take decades [19].

We aim to automate the measurement of bone lengths from a single, casually arranged, top-down image of a bird skeleton specimen. This casual capture (Figure 1) enables rapid digitization (taking 1 minute), but introduces challenges: bones intersect and appear in arbitrary rotations, and bone categories must be recognized across varying species. As a step towards to digitizing a larger ≈25K skeleton collection, we introduce an annotated dataset in Section 2 that includes both pixel annotations and hand-measurements. We describe a measurement system based on Mask-RCNN [8] and synthetic augmentation [6] in Section 3, which we train on our preliminary data.

Our experiments ( Section 4) show that our system, when trained on 150 preliminary specimens, automatically measures 60% of bones within 5% relative error. Some types (e.g., Humerus) are measured with 90% within 5% relative error and many bones are measured to sub-mm precision. We hope our method will enable large-scale ecological studies and contribute to our understanding of the links between climate change and ecology.

**Related Work:** Our work is part of a broader trend of using deep-learning based tools to solve research questions in the sciences. This has been applied to topics ranging from botany [15, 4, 18] to keypoint detection [14, 12] and or 3D reconstruction [3, 1, 2, 21]. While our animal specimens

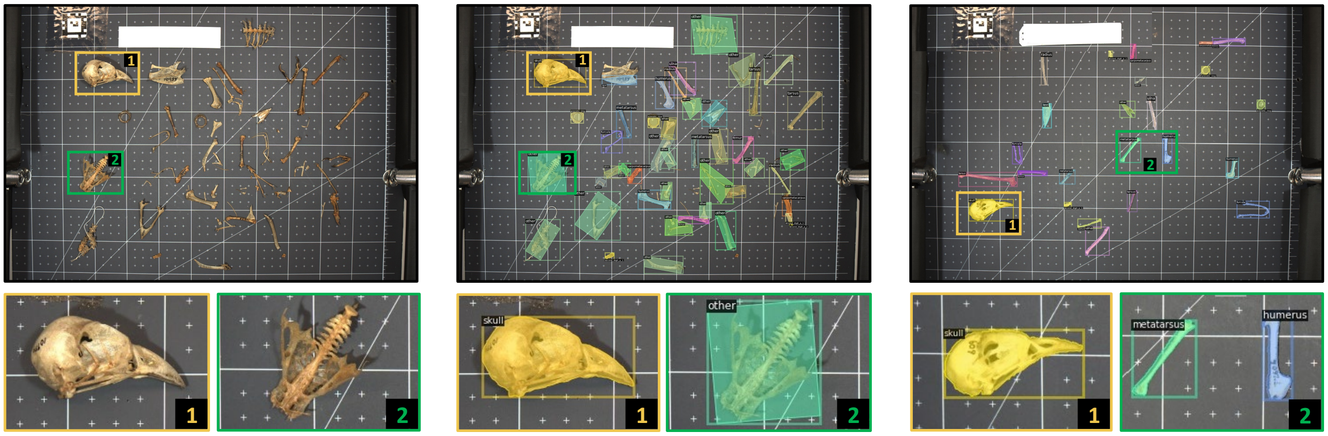

Figure 2. Visualization of original specimen image (left), annotated specimen image (middle), and synthetic specimen image (right).

do not move, making measurement easier, our measurements need high precision, which leads to a special capture setup. Our annotations require extensive expert time, which shapes data collection and leads to heterogeneously-annotated data, which we handle by self-training.

We frame our problem via instance segmentation to take advantage of the well-developed segmentation tools in the community [8, 20, 11]. In our work, we use Mask R-CNN [8] for our method, and the Toronto Annotation Suite [10] to accelerate our annotation. We similarly take inspiration from improvements to instance segmentation using synthetic data [6]. Building on this well-developed toolchain enables our approach to obtain promising performance with a fairly small amount of annotation.

## 2. Dataset

We evaluate methods on a 376 annotated image subset out of a larger in-progress effort that has captured over 10,000 bird skeleton images. Of the 376 images, 250 are annotated with pixel annotations. We hold out another 126 that are separately hand-measured with bone lengths; the non-unified set stems from Covid-19 restrictions. For now, we focus on 10 **Bones of Interest**, labeling other bones as **other**. We collect our dataset in four stages. All images are *Captured* in a process that is rapid and does not require specialized placement. Our annotated data is then processed with multiple stages to maximize use of experts' time, consisting of a *Bone Spotting* done by a trained technician, and *Pixel Annotation* that can often done by a non-trained expert. Finally, for empirical validation, we also do *Physical measurement* annotation of a small number of specimens.

**Image Capture:** Capturing a top-down image of a specimen entails retrieving the specimen and spreading the bones onto the capture surface. This is done casually to ensure the process takes only around a minute. These bones are imaged with a FLIR Blackfly S camera with a SONY IMX183 sensor ≈400mm above the center of the capture surface. In setup, 1px error corresponds to ≈0.07mm error.

**Bone Spotting:** Pixel-annotation of all bones of interest takes a long time, so to enable parallelizability and efficiently use expert time, we first complete a bone spotting step. In this step, the expert provides a rough bounding polygon and name of all bones. This enables parallelizing the slower segmentation task. We use the VGG Image Annotator [5] to annotate bounding polygons.

**Pixel Annotations:** Once bones of interest have been separated, we annotate cropped images around the bounding polygons. Easy bones (e.g., humerus) can be annotated by non-experts. We use the Toronto Annotation Suite [10], powered by human-in-the-loop AI, which substantially accelerates annotation. While all bones have a bounding polygon from the expert annotation stage, "other" and broken ones do not have pixel segmentations. This creates a mixed dataset of bounding polygons and pixel-wise segmentations, shown in Figure 2. We describe our approach to handle the mixed annotations in Section 3.

**Physical Measurement:** To provide end-to-end validation of the full system in terms of measurement in physical quantities, we hand-measure a held-out test set of 126 specimens. We hand measure the length of each bone of the specimen and record it to the nearest hundredth of a mm.

## 3. Method

We build a skeleton measurement system to compute physical measurements of the bones of interest from an image. Our core system is built on MaskRCNN [8]. We perform additional steps to enable best use of the data that is available: there are many "other" bones without pixel annotation, and the number of bones of interest with pixel segmentation is small. We produce a homogeneous dataset by using noisy self-training to annotate the other bones and synthetic data to expand our dataset.

Table 1. AP on a test set of 50 images with Mask R-CNN trained on different datasets described in Section 3.

| Dataset | mAP | Carpo-metacarpus | Femur | Humerus | Radius | Sclerotic Ring | Second Digit | Skull | Tarso-metatarsus | Tibio-tarsus | Ulna |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Original | 46.0 | 44.7 | 51.3 | 71.5 | 15.2 | 75.2 | 17.9 | 91.3 | 42.2 | 32.0 | 18.7 |
| Noisy | 52.2 | 47.7 | 55.4 | **72.5** | 18.4 | 76.2 | 23.1 | 92.1 | 41.1 | **57.5** | 37.6 |
| Synthetic | 38.1 | 25.9 | 41.7 | 61.9 | 7.6 | 64.0 | 8.0 | 81.6 | 18.9 | 38.8 | 32.4 |
| Noisy + Synthetic | **57.1** | **51.1** | **62.3** | 72.4 | **25.3** | **78.3** | **34.9** | **92.4** | **46.8** | 56.9 | **51.0** |

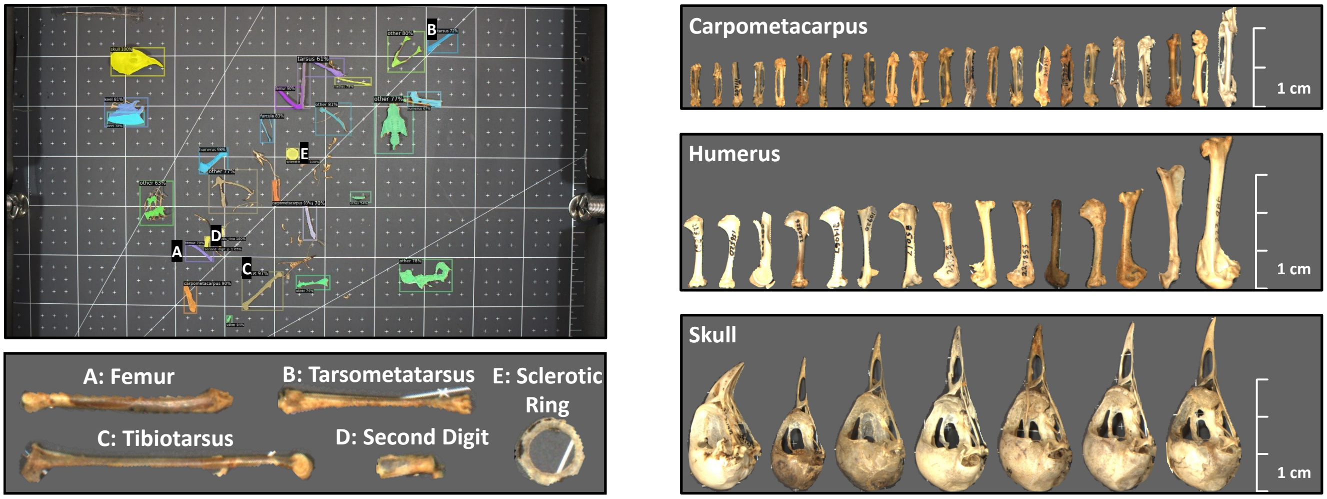

Figure 3. Left: a sample instance segmentation output with extracted and reoriented bones on the bottom. Right: a showcase for randomly selected carpometacarpus, humerus, and skull segmentations, demonstrating the potential of our system for large-scale skeleton analysis.

**Skeleton Measurement System:** The skeleton measurement system comprises an instance segmentation system which identifies the type of bone and its segmentation, followed by simple image processing to produce annotations. Specifically, we train Mask R-CNN [8] to predict bounding boxes plus class labels (the 10 bones of interest, plus other) as well as segmentation masks. Once these bones have been segmented, we extract the longest diagonal of the segmentation mask. This pixel measurement is converted to millimeters by assuming the bone is on the surface and that the surface and the image plane are parallel.

**Noisy Annotation Generation:** Since the “other bones” bones are spotted but not segmented, we create a homogeneous dataset by self-training. Given the initial annotations for the bones of interest, we train a binary U-Net [17] that predicts bone-vs-background; we minimize loss only on regions that have been pixel-annotated (i.e., ignoring other bones). The resulting annotation is then merged with the original dataset to produce what we call a **Noisy Dataset**.

**Synthetic Data Generation:** Finally, we expand our training set by building a synthetic dataset following [6]. This helps get additional performance out out of the limited annotations that can be generated. We generate images that mimics the data capture process by: (1) picking a background; (2) picking a random sample of segmented bones from the training set that models the skeletal composition of a bird; and (3) randomly selecting a location for each bone from a list of anchor locations. Finally, we composite the image with randomized blending parameters (Gaussian blur kernel size, and erosion kernel), similar to [6]. Parameter ranges were tuned manually to make the resulting images realistic. We refer to the results as the **Synthetic Dataset**.

**Implementation Details:** Our system uses Mask R-CNN [8] model with ResNet50 [9] and FPN [13] backbone using pre-trained weights from [20]. We train all models for 12,000 steps using Adam with base learning rate of .001 and the default Detectron2 [20] hyperparameters. We train with a resolution of 1368x912 pixels (4× smaller) for memory usage, making a pixel ≈0.28mm. Our self-training U-Net [17] has 4 max pooling and 4 up-sampling blocks, which each contain 2 convolutional layers with kernel size 3.

## 4. Experiments

We now evaluate the performance of our our bone measurement system. We evaluate the system in two modes – pixel-based metrics that test the system’s ability to segment bones of interest, and end-to-end metrics that test its ability to measure bones in millimeters. We show some examples of our system segmenting objects in Figure 3, including on a single image as well as showing a collection of segmented

Table 2. **(On All Specimens)** Percent of bones detected and measured within 5% relative error on a dataset of physically measured bones.

| Data | Percent<br>$|\epsilon| \leq 5\%$ | Carpo-<br>metacarpus | Femur | Humerus | Radius | Sclerotic<br>Ring | Second<br>Digit | Skull | Tarso-<br>metatarsus | Tibio-<br>tarsus | Ulna |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Original | 61.1 | 28.4 | 66.7 | 90.7 | 38.6 | 78.3 | 20.0 | **86.5** | **57.7** | 84.9 | 59.6 |
| Noisy | 63.8 | **40.9** | 78.4 | 87.2 | 42.0 | 69.6 | **25.9** | 84.6 | **57.7** | 89.5 | **61.8** |
| Noisy + Synthetic | **64.8** | **40.9** | **80.4** | **91.9** | **46.6** | **80.4** | 23.5 | 82.7 | 53.8 | **90.7** | 57.3 |

Table 3. **(On specimens found by all models)** Average error (mm) on a physically measured set of skeletons.

| Dataset | Mean<br>$\epsilon$ | Carpo-<br>metacarpus | Femur | Humerus | Radius | Sclerotic<br>Ring | Second<br>Digit | Skull | Tarso-<br>metatarsus | Tibio-<br>tarsus | Ulna |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Original | 1.10 | 1.38 | 0.59 | 0.86 | 2.36 | 0.28 | **0.56** | **1.43** | **1.16** | 1.24 | 1.12 |
| Noisy | 0.99 | 1.31 | 0.49 | 0.78 | **1.63** | 0.30 | 0.64 | 1.55 | 1.25 | **0.95** | 1.04 |
| Noisy + Synthetic | **0.92** | **1.08** | **0.43** | **0.52** | 1.65 | **0.27** | 0.60 | 1.49 | 1.26 | 1.04 | **0.89** |

Table 4. Mean Average Precision for different real/synthetic ratios by controlling the number of synthetic and real images.

| Data | 500 Syn. | 1k Syn. | 2k Syn. | 4k Syn. |
|---|---|---|---|---|
| 1x Real | **57.13** | 56.30 | 56.37 | 54.47 |
| 2x Real | 53.75 | 56.44 | 56.91 | 55.94 |
| 3x Real | 55.98 | **57.40** | **57.47** | 55.73 |
| 4x Real | 55.76 | 56.16 | 57.39 | **56.98** |

specimens.

**Experimental Setup and Metrics:** Throughout we split our segmented dataset of 250 images into 150 training images, 50 validation images, and 50 test images. We train and validate Mask-RCNN on the training and validation sets. When evaluating segmentation performance, we use the 50 test segmented images and quantify with Mean Average Precision (mAP) between IoU thresholds of 0.5 and 0.95. When evaluating measurement metrics, we use our 126 held-out hand-measured specimens for testing. Given an error $\epsilon$, we quantify performance by the percent of bones detected and measured within 5% relative error.

**Segmentation-based Evaluation:** We compare Mask R-CNN trained on the original, noisy, synthetic, and noisy + synthetic datasets in Table 1. Noisy data offers improvement over the original data while training on a both noisy and synthetic data offers drastic improvements over the original data. These improvements are particularly strong for poor-performing bones like the carpometacarpus, second digit, and ulna, but other better-performing bones (e.g., humerus) do not benefit or substantially degrade.

**Changing the Number of Synthetic Images:** Since synthetic data improves performance, we next explore how much to use. Table 4 shows that adding over 500 synthetic images decreases performance with just one real set. If we duplicate real data during training, peak performance is between 1k and 2k synthetic images. If there is too much synthetic data, performance decreases, likely due to overfitting to artifacts. Table 1 shows an extreme case where only synthetic data has poor performance.

**End-to-End Evaluation:** We finally evaluate the end-to-end performance by analyzing the percent of bones that are detected and measured within 5% relative error across different classes. Table 2 shows a similar trend: noisy and synthetic data aids performance. Gains are not as strong as in Table 1, suggesting that segmentation accuracy is correlated, but only partially so, with measurement performance.

To give a sense of the scale of bones, we also report the mean error in Table 3, computed only on bones detected by all three models in comparison. This comparison is needed to ensure apples-to-apples comparisons across models. Many average errors for bones are sub-millimeter (e.g., sclerotic ring), enabled by both highly accurate segmentation models and high-quality cameras. Even after scaling by the relatively small bones, these errors are small. For instance, the average humerus error (0.52mm) is 2.3% of the length of the average humerus (22.7mm). Again, increasing synthetic data improves results, but not as dramatically as in segmentation.

## 5. Conclusion

We presented a system for skeletal trait measurement. Our preliminary results suggest that the current pipeline can can obtain highly accurate results even with limited data. We expect that increasing the number of real images in use will substantially improve results. We believe our work will enable us to tap data on the relationship between climate and animal ecology on an unprecedented scale. Future work includes expanding bones of interest and increasing segmentation resolution.

**Acknowledgments:** This work is based on specimens from the University of Michigan Museum of Zoology, Department of Ornithology, with the support of Benjamin Winger, Curator of Birds, and Brett Benz, Collection Manager, Division of Ornithology.

## References

[1] Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd Pfrommer, Marc Schmidt, and Kostas Daniilidis. 3D bird reconstruction: a dataset, model, and shape recovery from a single view. In *ECCV*, 2020. 1

[2] Praneet C. Bala, Benjamin R. Eisenreich, Seng Bum Michael Yoo, Benjamin Y. Hayden, Hyun Soo Park, and Jan Zimmermann. Openmonkeystudio: Automated markerless pose estimation in freely moving macaques. *bioRxiv*, 2020. 1

[3] Benjamin Biggs, Ollie Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out: 3D animal reconstruction with expectation maximization in the loop. In *ECCV*, 2020. 1

[4] Kushtrim Bresilla, Giulio Demetrio Perulli, Alexandra Boini, Brunella Morandi, Luca Corelli Grappadelli, and Luigi Manfrini. Single-shot convolution neural networks for real-time fruit detection within the tree. *Frontiers in plant science*, 10:611, 2019. 1

[5] Abhishek Dutta, Ankush Gupta, and Andrew Zissermann. Vgg image annotator (via). *URL: http://www. robots. ox. ac. uk/˜ vgg/software/via*, 2016. 2

[6] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *ICCV*, pages 1301–1310, 2017. 1, 2, 3

[7] Janet L. Gardner, Anne Peters, Michael R. Kearney, Leo Joseph, and Robert Heinsohn. Declining body size: a third universal response to warming? *Trends in Ecology & Evolution*, 26(6):285–291, 2011. 1

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2961–2969, 2017. 1, 2, 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[10] Amlan Kar, Seung Wook Kim, Marko Boben, Jun Gao, Tianxing Li, Huan Ling, Zian Wang, and Sanja Fidler. Toronto annotation suite. `https://aidemos.cs.toronto.edu/toras`, 2021. 2

[11] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, pages 9799–9808, 2020. 2

[12] Siyuan Li, Semih Günel, Mirela Ostrek, Pavan Ramdya, Pascal Fua, and Helge Rhodin. Deformation-aware unpaired image translation for pose estimation on laboratory animals. In *CVPR*, pages 13158–13168, 2020. 1

[13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3

[14] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie W. Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 2018. 1

[15] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time blob-wise sugar beets vs weeds classification for monitoring fields using convolutional neural networks. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4, 2017. 1

[16] Alex L Pigot, Catherine Sheard, Eliot T Miller, Tom P Bregman, Benjamin G Freeman, Uri Roll, Nathalie Seddon, Christopher H Trisos, Brian C Weeks, and Joseph A Tobias. Macroevolutionary convergence connects morphological form to ecological function in birds. *Nature Ecology and Evolution*, 4:230–239 (2020), Jan. 2020. 1

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 3

[18] Yosuke Toda, Fumio Okura, Jun Ito, Satoshi Okada, Toshinori Kinoshita, Hiroyuki Tsuji, and Daisuke Saisho. Training instance segmentation neural network with synthetic datasets for crop seed phenotyping. *Communications biology*, 3(1):1–12, 2020. 1

[19] Brian C. Weeks, David E. Willard, Marketa Zimova, Aspen A. Ellis, Max L. Witynski, Mary Hennen, and Benjamin M. Winger. Shared morphological consequences of global warming in north american migratory birds. *Ecology Letters*, 23(2):316–325, 2020. 1

[20] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2 (2019). *URL https://github. com/facebookresearch/detectron2*. 2, 3

[21] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*, July 2017. 1