SPECIES-AWARE MORPHOMETRIC PRIORS FOR AVIAN 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing a bird in 3D from one photograph remains hard. There are 11,000 species of extreme variety, almost no 3D shape data, and no absolute scale in the image. Annotations mark one point per part, which says where a part is but not how long it is. Current methods regress the parameters of a bird model built from a few species, and they recover pose and silhouette well. Yet the part proportions they produce barely change from one species to another and do not match those measured for the species. Our starting observation is a difference between birds and humans. Humans remain one species however skin, height or limb length varies. Birds share the same joints, yet how far the bill, tarsus, tail and wing extend from those joints differs by species and is likely what tells one species from another. Ornithology measures exactly these lengths, and we bring them into 3D reconstruction from AVONET, an open database of more than 90,000 measured individuals across 11,000 species. Since a photograph has no scale, we follow the AVONET definitions and express each length as a ratio to wing length. Since these ratios vary little within a species, the species-average ratios serve as a species-aware morphometric prior. A morphometric pathway realizes it without disturbing the pretrained body model. A proportion head infers per-part length factors from the features of a pretrained transformer reconstructor, and an axial deformation operator inside the bird model applies them before skinning. Because the body model is left intact, pose and camera adapt to the new part lengths and species morphology is added without relearning the shape space. No photograph carries a measurement: for real images training uses species averages alone, and inference uses one photograph. On 49 held-out species the median error of the tail, bill and tarsus proportions against the measured species values falls from 21–31% without the prior to about 10%. On 237 species never seen in training it falls from 25–61% to 11–18%, while 2D keypoint accuracy improves and the 3D error is preserved. Long-legged species such as herons, cranes and storks, absent from the training photographs, remain difficult, and extending the prior to them is the next task.

1 INTRODUCTION

Reconstructing a bird in 3D from one photograph remains hard. About 11,000 bird species span an extreme range of forms (Tobias et al., 2022), yet 3D shape data barely exist. The parametric bird model in common use was itself built from photographs of 17 species (Wang et al., 2021). A photograph has no absolute scale, and the standard annotation marks one point per part (Wah et al., 2011), which says where a part is but not how long it is.

Current methods regress the parameters of such a model from the image, from the single-species model of Badger et al. (2020) to AVES (Wang et al., 2021) and the recent AniMer+ (An et al., 2026), which recovers pose and silhouette well. Measuring its reconstructions tells a different story. On the mesh in its rest pose we take tail, bill and tarsus length relative to wing length, for 7,408 photographs of 433 species. Across species the measured ratios vary by 35–43%, while the ratios on the reconstructed meshes vary by 2–6% (Section 4.2). Whatever the species, the reconstructed bird has almost the same part proportions. The shape space comes from a few species, keypoints and silhouettes carry no signal about length, and the standard metrics do not measure it. We show that this property can be supervised from species-level statistics alone, with no measurement on

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

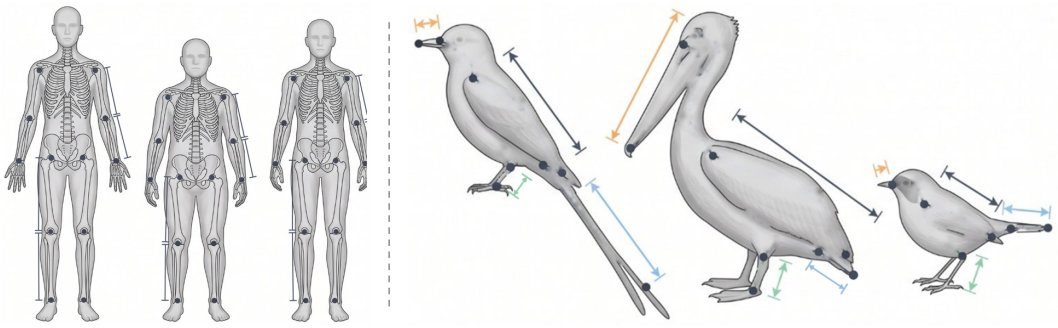


Figure 1: **Part proportions in humans and in birds**. Left: three humans of different build, with the same joints (dots) and nearly the same ratios between segment lengths (grey lines: upper arm, forearm, thigh, shin), all of one species. Right: three birds of different species, with the same joints but very different ratios of bill (orange), tarsus (green) and tail (blue) length to wing length (dark grey). The marked lengths are those measured in ornithology. Both panels are schematic.

any image. A reconstruction that is to serve measurement, biomechanics or biodiversity monitoring must carry the geometry of the species, not only its pose.

Our starting observation is a difference between birds and humans. Both have a fixed set of joints, and in both one can compare the ratio between two lengths of the skeleton. In humans such ratios, for example shin to thigh or forearm to upper arm, differ only slightly from one person to the next (Figure 1, left). Where they do differ (Ruff, 2002), the difference is between individuals of one species, and the shape space of a human body model describes exactly this variation (Loper et al., 2015). In birds the corresponding ratios, such as tarsus to wing or bill to wing, differ severalfold between species (Figure 1, right). These lengths are function, not ornament. The bill is the primary apparatus for capturing and processing food (Tobias et al., 2022), and legs (Zeffer et al., 2003) and tails (Thomas, 1993) differ with the way a bird moves. About 98% of the variance of such measurements lies among species, not within them (Pigot et al., 2020). What a body model treats as variation between individuals is, in birds, largely a property of the species, so it can be supervised from measurements of the species.

These lengths are measured in ornithology with standardized protocols, and AVONET (Tobias et al., 2022) provides them for more than 90,000 individuals of 11,000 species. Since a photograph has no scale, we follow the AVONET definitions and express each length as a ratio to wing length. These ratios vary little within a species, so the species-average ratios serve as a *species-aware morphometric prior*. A *morphometric pathway* realizes the prior without disturbing the pretrained body model. A *proportion head* infers per-part length factors from the features of the frozen AniMer+ backbone, and an axial deformation operator applies them inside the bird model before skinning. Body shape stays exactly as the pretrained model predicts it, so pose and camera adapt to the new part lengths. For real photographs the only supervision of proportions is the species average. No photograph carries a measurement, and inference needs neither a species label nor AVONET.

On 49 held-out species the median error of the three ratios falls from 21–31% to 9–11%. On 237 species never seen in training it falls from 25–61% to 11–18%. 2D keypoint accuracy improves, the 3D error is preserved, and the mesh remains well formed. Adding further species, which needs no annotation, lowers the error on species that are never shown. Our contributions are:

- A protocol that measures the part proportions of reconstructed birds against field measurements, revealing a gap that standard metrics miss.
- A species-aware morphometric prior that turns species-average measurements into a training signal.
- A morphometric pathway that realizes the prior without touching body shape, with an analysis of why three alternatives fall short.
- A three-seed evaluation on held-out and unseen species, and an annotation-free extension of the species range.

2 Related Work

3D reconstruction of animals and birds. SMAL (Zuffi et al., 2017) learned a quadruped shape space from 41 scans of toy figurines. Its parameters are now predicted from one image, as in SMALST (Zuffi et al., 2019) and WLDO (Biggs et al., 2020), with regressors of the kind used for human mesh recovery, HMR (Kanazawa et al., 2018a) and HMR 2.0 (Goel et al., 2023). Most of this line addresses quadrupeds, including the recent SAM 3D Animal (Hu et al., 2026) and AnimalLift (Sun et al., 2026). Badger et al. (2020) built the first parametric mesh from one species, AVES (Wang et al., 2021) extended it to a multi-species shape space, and AniMer+ (An et al., 2026), which extends the quadruped model AniMer (Lyu et al., 2025) to birds, scales this line with a large transformer and synthetic 3D data. Template-free methods learn category shape from image collections without 3D supervision (Appendix C). Length scales of parts already exist in these models, as limb scales in WLDO and as bone, bill and tail scales in the bird models. Their values come from image fitting, a generic prior or synthetic labels, and on real bird photographs AniMer+ constrains bone lengths only by a prior toward their mean. Evaluation uses 2D keypoints and silhouettes on photographs, and 3D error on synthetic animals or on scans of toys. In none of these works, for quadrupeds or for birds, are reconstructed part lengths compared with measurements of real animals.

Taxon information and measurements as shape supervision. The closest prior work is BARC (Rüegg et al., 2022), which uses dog breed labels during training only. It applies breed similarity losses to a latent code and pulls the shape coefficients and limb scales of 11 breeds toward computer-graphics models and toy figurines. It reports that its similarity loss does not work well on the shape coefficients directly, consistent with our ablation in Section 4.3. BITE (Rüegg et al., 2023) adds an evaluation on scans of real dogs. We differ in three respects. Our signal is a physical measurement of real animals, not a label of which animals look alike. The quantity we supervise lies outside the shape space of the bird model and needs a separate pathway. Our main evaluation compares part lengths with measurements. For humans, SHAPY (Choutas et al., 2022) supervises virtual measurements of the rest-pose mesh with anthropometric measurements of the photographed person, and our ruler follows this idea. Our targets are instead species averages, and we test on 237 unseen species rather than new individuals of one species.

Morphometrics from images. AVONET (Tobias et al., 2022) underlies global analyses that link avian form to ecological function (Pigot et al., 2020). Bone lengths have been measured automatically from photographs of museum skeletal specimens (Weeks et al., 2023), and the body length of whales from drone imagery (Gray et al., 2019). Most relevant is WildProp (Chasmai et al., 2026). It estimates the distribution of 2D part-length ratios of a species from one annotated query image and a collection of citizen-science photographs, with relative errors of 10–20% against physical measurements. We use the same measurements and the same definition of error. The task differs. WildProp estimates 2D ratios from an image collection, while we make the 3D mesh reconstructed from one photograph carry those ratios. Appendix G compares the two on the six species they share.

3 Method

We keep a pretrained reconstructor and its body model fixed (Figure 2), define what is measured on the mesh (Section 3.2), add a pathway that changes only those quantities (Section 3.3) and train it from species averages (Section 3.4).

3.1 Preliminaries

AVES (Wang et al., 2021) generates a bird mesh from a template $\bar{T} \in \mathbb{R}^{N\times 3}$ and a linear shape basis B with coefficients $\boldsymbol{\beta}$. Bone length scales $\boldsymbol{\alpha}$ stretch the distances between joints, and $\boldsymbol{\theta}$ collects the joint rotations. The shaped template is $T(\boldsymbol{\beta}) = \bar{T} + B(\boldsymbol{\beta})$ and the mesh is

$$V(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = W\big(T(\boldsymbol{\beta}), \boldsymbol{\alpha}, \boldsymbol{\theta}\big), \tag{1}$$

where W is linear blend skinning. We call $\hat{V} = W(T(\boldsymbol{\beta}), \boldsymbol{\alpha}, \mathbf{I})$, the mesh with all joint rotations set to the identity, the *rest mesh*. AniMer+ (An et al., 2026) computes features $F = \Phi(I)$ of an image I with a ViT (Dosovitskiy et al., 2021) backbone Φ as in ViTPose++ (Xu et al., 2024), and a head H regresses $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ from F, where $\boldsymbol{\gamma}$ is the camera. 2D keypoints are $x = \Pi(KV; \boldsymbol{\gamma})$ with a fixed

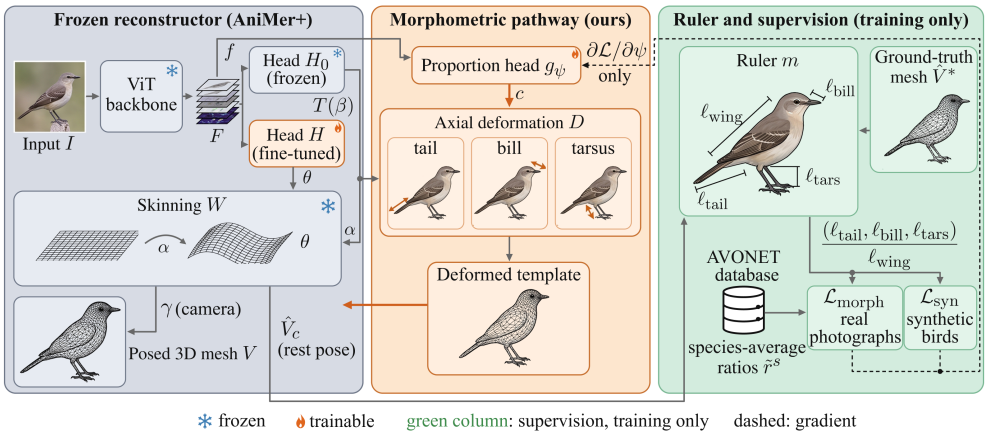


Figure 2: **Overview of the morphometric pathway**. The backbone and the predicted body shape are frozen (AniMer+ in our experiments); the head H that gives pose and camera is fine-tuned. The proportion head predicts three length factors, the axial deformation operator applies them to the shaped template $T(\boldsymbol{\beta})$ before skinning, giving the deformed template, and the ruler measures the resulting part-to-wing ratios. On real photographs the ratios are compared with the species average from AVONET, with the tail target scaled by the ruler constant k, and on synthetic birds with the ratios of the ground-truth mesh. The ruler measures $\hat{V}_c$, the rest mesh output by the skinning. g_ψ and D are ours. Losses on keypoints, pose and rehearsal are omitted.

keypoint regressor K and perspective projection Π. We write H_0 for a frozen copy of the pretrained head. Appendix A lists the notation.

3.2 Species-aware morphometric prior

A ruler on the mesh. The ruler m maps a rest mesh to three ratios,

$$m(\hat{V}) = \Big(\frac{\ell_{\text{tail}}(\hat{V})}{\ell_{\text{wing}}(\hat{V})}, \frac{\ell_{\text{bill}}(\hat{V})}{\ell_{\text{wing}}(\hat{V})}, \frac{\ell_{\text{tars}}(\hat{V})}{\ell_{\text{wing}}(\hat{V})} \Big), \tag{2}$$

Each length transfers the measurement definition of AVONET (Tobias et al., 2022) to the mesh: tail to the tip of the longest feather, bill along the culmen, tarsus between its two joints and wing along the folded wing (Appendices B and D). Measuring the predicted mesh in a rest pose with a differentiable function follows the virtual measurements of SHAPY (Choutas et al., 2022).

The prior. For a species s with mean lengths $\bar{\ell}^s$ over the individuals measured in AVONET, the prior is

$$\bar{r}^s = \Big(\frac{\bar{\ell}^s_{\text{tail}}}{\bar{\ell}^s_{\text{wing}}}, \frac{\bar{\ell}^s_{\text{bill}}}{\bar{\ell}^s_{\text{wing}}}, \frac{\bar{\ell}^s_{\text{tars}}}{\bar{\ell}^s_{\text{wing}}} \Big). \tag{3}$$

We use ratios because a photograph has no absolute scale. We use species averages because the ratios are a property of the species. On the 54,219 individuals of 6,586 species with all four measurements and at least five individuals, the share of the variance of the log ratios that lies among species is 0.93, 0.94 and 0.95 for tail, bill and tarsus. Within a species the ratios vary by 6.0–6.5% (median coefficient of variation). The ratios alone do not identify a species (Appendix B). For the tail, the ruler on the mesh and the measurement in AVONET do not span the same interval. We therefore multiply the tail target by one global constant, $\tilde{r}^s_{\text{tail}} = k\,\bar{r}^s_{\text{tail}}$, and keep $\tilde{r}^s_p = \bar{r}^s_p$ for the other parts. The constant k is estimated once, before training, from training photographs (Section 3.5).

3.3 Morphometric pathway

Proportion head. The proportion head g_ψ takes the vector f, which concatenates the mean of the patch tokens and the class token of F. It predicts one length factor per part,

$$c = g_\psi(f) = l + \sigma\big(\text{MLP}_\psi(f) + u_0\big) \odot (u - l), \tag{4}$$

where σ is the logistic function, l and u are lower and upper bounds for each part, and u_0 is the logit at which c equals the neutral factor c_0 that leaves the mesh unchanged. The last layer of the MLP is initialized at zero, as are the branches added by LoRA (Hu et al., 2022) and ControlNet (Zhang et al., 2023). Training therefore starts at $c = c_0$, that is, exactly at the pretrained reconstructor.

Axial deformation operator. For each part p the operator D uses a root point o_p and a unit axis a_p, read from the mesh and its rest joints. It also uses a part weight $w_p(v)$ derived from the skinning weights of AVES (Wang et al., 2021) (the summed weights of the part for bill and tarsus, and an indicator of the tail vertices for the tail), and an axial profile ρ_p. It moves every vertex v to

$$v' = v + s_p(c_p)\,\rho_p\big(\langle v - o_p,\, a_p\rangle\big)\,w_p(v)\,a_p, \tag{5}$$

where s_p maps the factor to a gain that vanishes at c_0. The profile ρ_p is zero at and behind the root and does not decrease along the axis. The displacement is thus zero at the joint, largest at the tip of the part, and zero on the body, where $w_p = 0$. It is purely axial, so the thickness of the part is preserved. The upper and lower bill move together, and both legs share one factor. Skinning at the identity pose is a translation of each vertex, so the displacement is added to the shaped template $T(\boldsymbol{\beta})$ and appears unchanged on the rest mesh. Under a pose it rotates with the blended rotation of the surrounding vertices. We write $\hat{V}_c = W\big(D(T(\boldsymbol{\beta}), c), \boldsymbol{\alpha}, \mathbf{I}\big)$ for the deformed rest mesh; by the argument above it equals $D(\hat{V}, c)$. The template, the shape basis and the skinning weights of AVES are not modified, and the body shape $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is taken from H_0 without change.

Why a separate pathway. The supervision has to act on part lengths directly, since keypoints and silhouettes do not constrain them, and it must not reach the parameters of body shape. Length scales exist in earlier models, as limb scales in WLDO (Biggs et al., 2020) and BARC (Rüegg et al., 2022), bone scales in the bird model of Badger et al. (2020) and bill and tail scales in AVES (Wang et al., 2021). There no measurement of the species constrains them. Section 4.3 tests three alternatives under the same protocol: the same loss on $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, 2D keypoints alone, and stretching the mesh after reconstruction.

3.4 Training objective

The objective is

$$\mathcal{L} = \mathcal{L}_{\text{2D}} + \lambda_m \mathcal{L}_{\text{morph}} + \lambda_s \mathcal{L}_{\text{syn}} + \lambda_\theta \mathcal{L}_{\text{pose}} + \lambda_r \mathcal{L}_{\text{reh}}. \tag{6}$$

Morphometric loss. On a real photograph I_i of species s_i, with body shape $(\boldsymbol{\beta}_i, \boldsymbol{\alpha}_i) = H_0(F_i)$ and factors $c_i = g_\psi(f_i)$, the loss compares the ratios measured on the deformed rest mesh with the prior,

$$\mathcal{L}_{\text{morph}} = \frac{1}{3|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{p} \Big| \log m_p\big(D(\hat{V}(\boldsymbol{\beta}_i, \boldsymbol{\alpha}_i), c_i)\big) - \log \tilde{r}_p^{\,s_i} \Big|, \tag{7}$$

where $\mathcal{B}$ is the batch and p runs over the three parts. No image carries a measurement. The species label is used only to look up the prior, and only during training. Because $(\boldsymbol{\beta}_i, \boldsymbol{\alpha}_i)$ come from the frozen head, the gradient of this loss reaches ψ and nothing else.

Synthetic ratio loss. On a synthetic bird I_j with ground-truth parameters $(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}_j^*)$, the target is the ratio of the ground-truth rest mesh,

$$\mathcal{L}_{\text{syn}} = \frac{1}{3|\mathcal{B}'|} \sum_{j \in \mathcal{B}'} \sum_{p} \Big| \log m_p\big(D(\hat{V}(\boldsymbol{\beta}_j, \boldsymbol{\alpha}_j), c_j)\big) - \log m_p\big(\hat{V}(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}_j^*)\big) \Big|. \tag{8}$$

Both sides use the same ruler, so no constant k is involved. Synthetic birds have shapes near those of the species used to build the bird model, so the prior of Equation 3 is not the right target for them. Without this term the proportion head never sees a synthetic bird during training, and at test time it deforms synthetic birds toward species averages that they do not follow.

Remaining terms. $\mathcal{L}_{\text{2D}}$ is the confidence-weighted L_1 distance between projected and annotated keypoints, including a tail-tip point that we extract from the mask. $\mathcal{L}_{\text{pose}}$ is the squared distance between the predicted joint rotations and those of H_0. $\mathcal{L}_{\text{reh}}$ is the loss of AniMer+ (An et al., 2026) on its synthetic 3D data without the silhouette term, used as rehearsal. The trainable parameters are ψ and the head H, of which only the pose and camera outputs are used. Since the projected keypoints come from the deformed mesh, $\mathcal{L}_{\text{2D}}$ updates ψ as well as H. At inference the model takes one photograph and uses neither a species label nor AVONET (Tobias et al., 2022).

Table 1: **Proportion error and correlation against AVONET**. Median relative error over species (%) and Pearson correlation r between predicted and measured species ratios, for tail, bill and tarsus length relative to wing length. CUB-49: species held out from supervision. NABirds-237: species seen by no component. Ours is the mean of three seeds, with the standard deviation below.

	CUB-49						NABirds-237					
	Error (%) ↓			r ↑			Error (%) ↓			r ↑		
Method	Tail	Bill	Tars.	Tail	Bill	Tars.	Tail	Bill	Tars.	Tail	Bill	Tars.
AniMer+	31.3	21.0	27.9	−0.32	0.22	−0.17	60.8	32.6	25.2	−0.46	0.50	0.30
Ours	**8.9**	**11.1**	**11.0**	**0.87**	**0.79**	**0.90**	**11.2**	**18.2**	**12.7**	**0.89**	**0.70**	**0.63**
s.d.	0.6	1.1	1.1	0.02	0.00	0.01	0.9	0.1	0.3	0.01	0.03	0.00

3.5 Implementation details

We set $\lambda_m = \lambda_s = 0.5$, $\lambda_\theta = 1$ and $\lambda_r = 0.1$ and train for 4,500 steps with 12 photographs and 12 synthetic images per step, about one hour on a single RTX 4090. The loss weights and learning rates were set once and not tuned. The number of steps was checked on a validation set of unused photographs of the training species. The constant $k = 0.445$ is estimated once from the training photographs with the pretrained reconstructor held fixed (Appendix F). Architecture, bounds and optimizer are given in Appendix E.

4 Experiments

We test whether the prior corrects the proportions without harming pose, 3D error or the mesh (Section 4.2), whether a separate pathway is needed (Section 4.3), and what is learned for unseen species (Sections 4.4 and 4.5).

4.1 Setup

Data. We split the 200 species of CUB (Wah et al., 2011) by species. 150 species with 4,075 photographs are used for training. 2,004 further photographs of the same species form the validation set. The remaining 50 species, with 678 photographs, are held out from all supervision. Species statistics use the 49 held-out species that have at least five photographs. To test species that no component has seen, we take 237 species of NABirds (Van Horn et al., 2015). They occur neither in CUB nor among the species used to build AVES (Wang et al., 2021), and each has 20 photographs (4,740 in total). The synthetic 3D set of AniMer+ (An et al., 2026) provides 6,464 training images, used for $\mathcal{L}_{\text{reh}}$ and $\mathcal{L}_{\text{syn}}$, and 309 test images with ground-truth meshes.

Metrics. *Proportions.* We measure the three ratios on each reconstructed rest mesh with the ruler of Section 3.2 and take the median over the photographs of a species. The *proportion error* of a part is the median over species of the relative error against the target $\tilde{r}^{\,s}$ of Section 3.2. The Pearson correlation r between predicted and measured species ratios does not depend on k. We compare correlations with a bootstrap over species (2,000 resamples). *Standard metrics.* We report PCK@0.1 on the CUB test set, the PCK of the tail tip, which CUB does not annotate, and PA-MPJPE and PA-MPVPE on the synthetic test set. *Mesh.* A model passes the *mesh gate* if no triangle shrinks below 10% of its area, no bill edge at the skinning boundary stretches beyond 1.5 (99th percentile) and no bone scale is negative (Appendix G). Results are means over three seeds, the number of training steps is chosen on the validation set, and the thresholds of every check are listed in Table 8.

4.2 Main results

The gap. We first measure the reconstructions of AniMer+ (An et al., 2026) on the 433 species with at least five photographs, 196 of CUB (Wah et al., 2011) and 237 of NABirds (Van Horn et al., 2015) (7,408 photographs). Across these species the coefficient of variation of the measured ratios is 38.9, 42.8 and 35.3% for tail, bill and tarsus. For the ratios on the reconstructed meshes it is 5.6, 2.0 and 2.4%, even though AniMer+ was trained on photographs of all CUB species. The reconstructed proportions hardly depend on the species (Figure 3, grey circles).

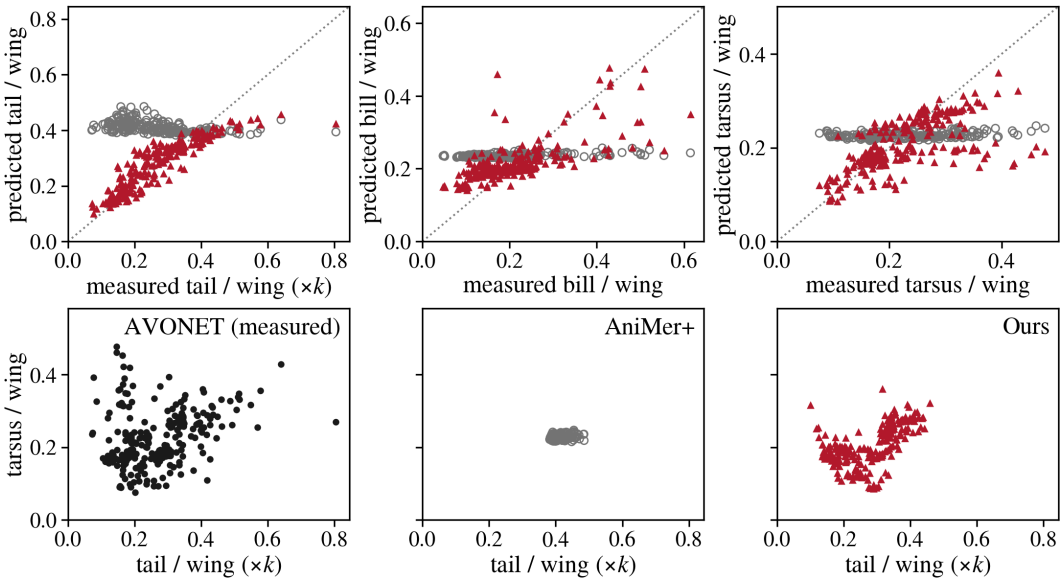


Figure 3: **Part proportions of reconstructed birds against field measurements**. The 237 species seen by neither model in training, one marker per species (black dots: AVONET, grey circles: AniMer+, red triangles: ours, mean of three seeds). Top: tail, bill and tarsus length relative to wing length, measured in AVONET (horizontal) against measured on the reconstructed mesh (vertical). The dotted diagonal marks perfect agreement. Bottom: the species in the plane of tail and tarsus ratios, as measured and as reconstructed. The measured tail ratio is multiplied by the ruler constant k of Section 3. Correlations are given in Table 1.

Proportions. Table 1 shows that species averages alone are enough to make the reconstructed proportions follow the species. On the held-out species the proportion error falls from 31.3, 21.0 and 27.9% to 8.9, 11.1 and 11.0% for tail, bill and tarsus. The ratios of AniMer+ are weakly or negatively correlated with the measurements, while ours reach r of 0.87, 0.79 and 0.90. The gain in r is significant for every part and every seed ($p < 0.001$). The spread among species is also recovered, with a standard deviation of 66, 74 and 79% of that of the measured ratios, against 21, 6 and 7% for AniMer+. Figure 4 shows four of these held-out species.

Unseen species. The same holds for species that no component has seen (Table 1, right, and Figure 3). On NABirds-237 the error falls from 60.8, 32.6 and 25.2% to 11.2, 18.2 and 12.7%. The largest bootstrap p over the three seeds is 0.0005 for tail, 0.002 for tarsus and 0.018 for bill. The bill gains least because AniMer+ already orders bill ratios partly ($r = 0.50$). Our ratio is the closer one for 208, 181 and 163 of the 237 species (five in Appendix G). In the plane of tail and tarsus ratios (Figure 3, bottom), our species cover 32.6% of the hull area of the measured species, those of AniMer+ 1.3% (Appendix H).

Standard metrics. Table 2 shows that the proportions are not bought with the standard metrics. 2D accuracy improves, most for the tail tip, because a tail of the right length places its tip where the photograph shows it. On the NABirds-237 photographs PCK@0.1 rises from 0.761 to 0.783 (Appendix H). The 3D error is preserved and the mesh passes the gate. The silhouette IoU drops slightly, since our training uses no silhouette term.

4.3 Alternatives to the pathway

Table 3 compares the pathway (Row 2) with the three alternatives of Section 3.3 (Rows 3–6). A usable model has to pass four checks, fixed before the runs relative to our model. *Ratios*: error at most 2 points above ours on CUB-49 (Wah et al., 2011) and NABirds-237 (Van Horn et al., 2015). *2D*: PCK at most 0.02 below. *3D*: PA-MPVPE at most 5% above. *Mesh*: the gate of Section 4.1. Each alternative matches or beats our model on one number but fails at least one check. Only the pathway passes all four.

Table 2: **All metrics for the base model and ours**. Error: proportion error averaged over tail, bill and tarsus (%) on CUB-49 and NABirds-237 (NAB-237), the means of Table 1. PCK: PCK@0.1 on the official CUB test set. Tip: PCK of the tail tip on CUB-49. PA-J and PA-V: PA-MPJPE and PA-MPVPE on the synthetic test set. IoU: silhouette IoU on CUB-49. Area: smallest triangle area ratio of the deformed mesh. Ours is the mean of three seeds. Best value per column in bold.

	Error (%) ↓		2D ↑		3D ↓			
Method	CUB-49	NAB-237	PCK	Tip	PA-J	PA-V	IoU ↑	Area ↑
AniMer+	26.7	39.5	0.594	14.5	**655.7**	709.0	**0.605**	**1.00**
Ours	**10.3**	**14.0**	**0.643**	**30.2**	658.9	**707.5**	0.597	0.64

Table 3: **Alternatives to the pathway**. Left: proportion error on CUB-49 (%), PCK@0.1, PA-MPVPE on synthetic birds (PA-V) and smallest area ratio of the deformed mesh (Area), means over three seeds. Right: checks of Section 4.3 relative to our model (✓ passed, × failed). All values are in Appendix Table 6.

		Measured				Checks			
#	Method	Error ↓	PCK ↑	PA-V ↓	Area ↑	Ratios	2D	3D	Mesh
1	AniMer+	26.7	0.594	709.0	1.00		reference		
2	**Ours**	10.3	0.643	707.5	0.64	✓	✓	✓	✓
3	Ratio loss on $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, anchored	24.9	0.632	704.0	0.38	×	✓	✓	✓
4	Ratio loss on $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, free	18.0	0.667	719.8	0.04	×	✓	✓	×
5	2D keypoints only	119.1	0.655	706.1	0.40	×	✓	✓	×
6	Stretching after reconstruction	9.1	0.594	827.5	0.59	✓	×	×	✓

Supervising the shape parameters. Rows 3–4 apply the same ratio loss to $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and use no pathway. Anchored to the pretrained prediction by a quadratic penalty (Row 3), the proportions hardly move (24.9% against 26.7% in Row 1). Free (Row 4), the error falls to 18.0%, still 7.7 points above Row 2, and the mesh breaks: the smallest area ratio is 0.04 and bone scales turn negative. Its PCK is higher than ours, so 2D accuracy can be bought with a damaged mesh.

2D keypoints alone. Row 5 trains the pathway without the ratio losses, so the factors are constrained by keypoints only. PCK is higher than in Row 2, but bill and tarsus ratios become meaningless and the bill stretches at its skinning boundary, so the mesh leaves the gate.

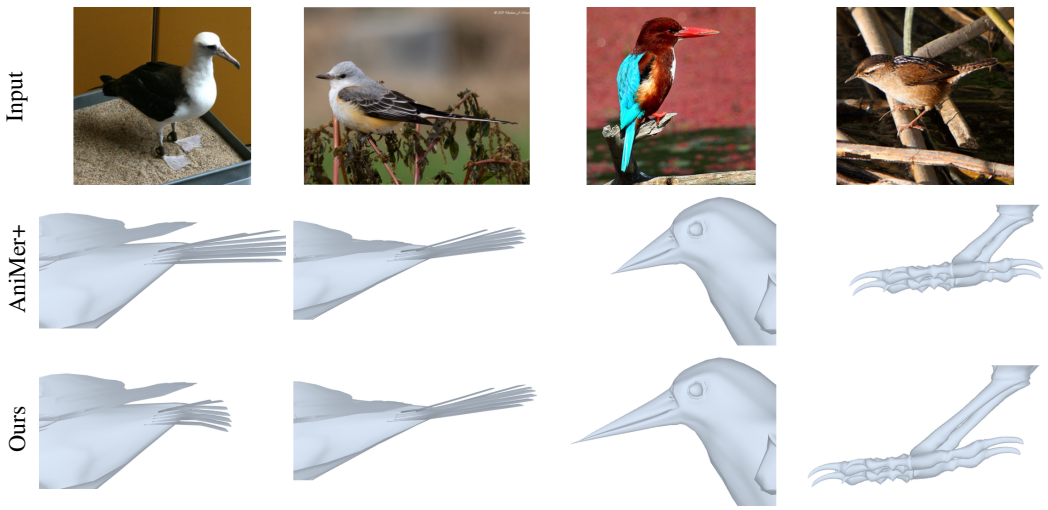
Stretching after reconstruction. In Row 6 a predictor trained on the same frozen features outputs the three factors, and the AniMer+ mesh is stretched with them after reconstruction. Its ratios on CUB-49 are as good as those of Row 2 (9.1%) and the mesh stays valid. Pose and camera, however, were estimated for the old part lengths. PCK stays at the level of Row 1, the PCK of the tail tip is 20.0 against our 30.2, and PA-MPVPE rises to 827.5. Putting the pathway inside the reconstructor is what lets the rest of the prediction adapt.

Components. Without $\mathcal{L}_{\text{syn}}$ the 3D error rises, a free body shape breaks the mesh and without k the tail error rises (Appendix G).

4.4 What is learned for unseen species

Not a lookup of the nearest species. A baseline returns the ratios of the most similar training species, by cosine similarity of the mean frozen features. Its correlations on NABirds-237 (Van Horn et al., 2015) are 0.71, 0.39 and 0.42, against 0.89, 0.70 and 0.63 for our model.

Extreme proportions are compressed. Regressing our species ratios on the measured ones over the unseen species gives slopes of 0.68, 0.41 and 0.43. The order of species is recovered, but long parts come out too short and short parts too long. It is strongest for nine long-legged waterbirds such as herons, with a median tarsus ratio of 0.42 measured and 0.19 ours. The factor is not saturated (at most 1.15, against a bound of 4), and four of the nine species lie inside the range of the training targets. The likely cause is the coverage of training (Section 4.5): among the training species a high tarsus ratio comes from a short wing, not from a long leg. The features therefore never pair a visibly long leg with a large factor.



(a) Short tail (b) Long tail (c) Long bill (d) Long tarsus

Figure 4: **Reconstructed meshes of four held-out species**. (a) Laysan Albatross (short tail), (b) Scissor-tailed Flycatcher (long tail), (c) White-breasted Kingfisher (long bill) and (d) Marsh Wren (long tarsus), chosen by the rule of Appendix G. Rows: input photograph, AniMer+ and ours. Each panel is a crop of the rest mesh at equal wing length, with the same box for both methods. The part-to-wing ratio measured in AVONET, on the AniMer+ mesh and on our mesh is 0.14 / 0.38 / 0.14 for the tail of the albatross, 0.79 / 0.39 / 0.44 for the tail of the flycatcher, 0.52 / 0.24 / 0.36 for the bill of the kingfisher and 0.40 / 0.23 / 0.34 for the tarsus of the wren.

4.5 Widening the covered species without annotation

The prior needs no annotation, so species can be added with photographs and AVONET (Tobias et al., 2022) alone. We split NABirds-237 (Van Horn et al., 2015) once, stratified by family, into 97 species that are never shown and a pool of 140. We add 35, 70 or all 140 pool species to training with at most 40 photographs each, without keypoints or masks, so that they contribute only $\mathcal{L}_{\text{morph}}$ (three seeds per setting). The gain is largest where the compression is strongest. The tarsus error of the 14 long-legged species among the never-shown ones falls from 28.5 to 8.3%. Over all never-shown species the tarsus slope rises from 0.43 to 0.66 and the mean proportion error goes from 12.34 to 12.31, 11.97 and 11.66%, while the bill does not improve (16.6 to 18.0%). The error on the validation set of the original species rises from 6.25 to 7.83%, CUB-49 (Wah et al., 2011) does not degrade (10.3 to 9.3%) and the mesh gate is passed throughout.

5 Conclusion and Limitations

What varies between individuals in humans varies between species in birds. We used this to bring the measurements of ornithology into 3D reconstruction: species-average ratios from AVONET (Tobias et al., 2022) serve as a prior, realized by an axial deformation before skinning that leaves the pretrained body model untouched. With no measurement attached to any photograph, the mean proportion error falls from 26.7 to 10.3% on held-out species and from 39.5 to 14.0% on unseen species. 2D accuracy improves and the 3D error is preserved. None of the three alternatives achieves all of this.

Limitations. The pathway leaves the shape space untouched, so everything outside the three lengths of tail, bill and tarsus remains as the base model predicts it. Since the tibia is not measured, a long-legged bird keeps short legs even with a correct tarsus ratio. The targets are species averages, so variation within a species is not modelled. On unseen species extreme proportions are compressed. Adding species eases it, at a small cost on the species already covered (Section 4.5). A strongly shortened tail can have a few flipped triangles (Appendix G). We evaluate one base reconstructor and one taxon. Skeletal traits (Weeks et al., 2025) and trait databases of other taxa could serve as similar priors.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on the experiments. We have not used generative AI tools for the other tasks with required disclosure, such as proposing hypotheses, designing the experiments or interpreting results. Additionally, we used generative AI tools to edit the wording of text written by the authors, to search for related work and to make a first check of the citations, which the authors then verified against the sources, to write and edit code for evaluation, analysis and figures, and to run experiments under settings fixed by the authors. We have reviewed all AI-assisted work and take responsibility for the final content of this work.

ETHICS STATEMENT

This work uses only publicly available datasets: CUB-200-2011, NABirds, the CtrlAVES3D set released with AniMer+, and AVONET, which is distributed under a CC-BY license. No human subjects are involved. NABirds photographs are shown with the names of their photographers, as the dataset asks.

REPRODUCIBILITY STATEMENT

The model, training recipe and evaluation protocol are described in Section 3 and Section 4.1, with the ratio definitions in Section 3.2 and the loss and all weights in Section 3.4. Implementation details are in Appendix E, the rulers and the deformation in Appendix D, the tail constant k in Appendix F, and the ablation criteria, fixed before the runs, together with all per-part values and seed deviations in Appendix G. Every reported number is produced by a script from a result file that carries a stamp of the training and evaluation settings, and the aggregation refuses to combine runs whose settings differ. Code, the species split, the AVONET aggregation and trained checkpoints will be released upon acceptance.

REFERENCES

Liang An, Jin Lyu, Li Lin, Pujin Cheng, Yebin Liu, and Xiaoying Tang. AniMer+: Unified pose and shape estimation across mammalia and aves via family-aware transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3233–3249, 2026. doi: 10.1109/TPAMI.2025.3633828.

Mehmet Aygün and Oisin Mac Aodha. SAOR: Single-view articulated object reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10382–10391, 2024.

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 (ECCV)*, pp. 1–17, 2020.

Chhandak Bagchi, Josh Medina, Duncan J. Irschick, Subhransu Maji, and Fredrik Christiansen. Automated extraction of right whale morphometric data from drone aerial photographs. *Remote Sensing in Ecology and Conservation*, 11(4):438–453, 2025.

Kevin Charles Bierlich, Sagar Karki, Clara N. Bird, Alan Fern, and Leigh G. Torres. Automated body length and body condition measurements of whales from drone videos for rapid assessment of population health. *Marine Mammal Science*, 40(4):e13137, 2024.

Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and SMAL: Recovering the shape and motion of animals from video. In *Asian Conference on Computer Vision (ACCV)*, pp. 3–19, 2018.

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 (ECCV)*, pp. 195–211, 2020.

Mustafa Chasmai, Aaron Sun, and Subhransu Maji. WildProp: Visual estimation of wildlife body proportions at scale. In *European Conference on Computer Vision (ECCV)*, pp. 430–449, 2026. doi: 10.1007/978-3-032-37013-6_23.

Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3D body shape regression using metric and semantic attributes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2708–2718, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *European Conference on Computer Vision (ECCV)*, pp. 88–104, 2020.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14737–14748, 2023.

Patrick C. Gray, Kevin C. Bierlich, Sydney A. Mantell, Ari S. Friedlaender, Jeremy A. Goldbogen, and David W. Johnston. Drones and convolutional neural networks facilitate automated and accurate cetacean species identification and photogrammetry. *Methods in Ecology and Evolution*, 10(9):1490–1500, 2019.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning (ICML)*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Xuyi Hu, Jin Lyu, Jiuming Liu, Yebin Liu, Silvia Zuffi, Liang An, and Stefan Goetz. SAM 3D Animal: Promptable animal 3D reconstruction from images in the wild. *arXiv preprint arXiv:2605.07604*, 2026.

Tomas Jakab, Ruining Li, Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Farm3D: Learning articulated 3D animals by distilling 2D diffusion. In *International Conference on 3D Vision (3DV)*, pp. 852–861, 2024.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7122–7131, 2018a.

Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *European Conference on Computer Vision (ECCV)*, pp. 386–402, 2018b.

Bernd Leisler and Hans Winkler. Ecomorphology. In *Current Ornithology*, volume 2, pp. 155–186. Plenum Press, 1985.

Ci Li, Yi Yang, Zehang Weng, Elin Hernlund, Silvia Zuffi, and Hedvig Kjellström. Dessie: Disentanglement for articulated 3D horse shape and pose estimation from images. In *Asian Conference on Computer Vision (ACCV)*, pp. 268–288, 2024a.

Ziqi Li, Abderraouf Amrani, Shri Rai, and Hamid Laga. Advances and trends in the 3D reconstruction of the shape and motion of animals. *arXiv preprint arXiv:2508.16062*, 2025.

Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3D fauna of the web. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9752–9762, 2024b.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Jin Lyu, Tianyi Zhu, Yi Gu, Li Lin, Pujin Cheng, Yebin Liu, Xiaoying Tang, and Liang An. AniMer: Animal pose and shape estimation using family aware transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17486–17496, 2025.

Seyed Mojtaba Marvasti-Zadeh, Mohammad N. S. Jahromi, Javad Khaghani, Devin Goodsman, Nilanjan Ray, and Nadir Erbilgin. Learning-based monocular 3D reconstruction of birds: A contemporary survey. *arXiv preprint arXiv:2207.04512*, 2022.

Donald B. Miles and Robert E. Ricklefs. The correlation between ecology and morphology in deciduous forest passerine birds. *Ecology*, 65(5):1629–1640, 1984. doi: 10.2307/1939141.

David R. Musicant, Janara M. Christensen, and Jamie F. Olson. Supervised learning by training on aggregate outputs. In *IEEE International Conference on Data Mining (ICDM)*, pp. 252–261, 2007.

Tomasz Niewiadomski, Anastasios Yiannakidis, Hanz Cuevas-Velasquez, Soubhik Sanyal, Michael J. Black, Silvia Zuffi, and Peter Kulits. Generative zoo. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8492–8502, 2025.

Giorgio Patrini, Richard Nock, Paul Rivera, and Tiberio Caetano. (Almost) no label no cry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

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 & Evolution*, 4:230–239, 2020.

Novi Quadrianto, Alex J. Smola, Tibério S. Caetano, and Quoc V. Le. Estimating labels from label proportions. In *International Conference on Machine Learning (ICML)*, pp. 776–783, 2008. doi: 10.1145/1390156.1390254.

Robert E. Ricklefs and Joseph Travis. A morphological approach to the study of avian community organization. *The Auk*, 97:321–338, 1980.

Nadine Rüegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. BARC: Learning to regress 3D dog shape from images by exploiting breed information. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3866–3874, 2022.

Nadine Rüegg, Shashank Tripathi, Konrad Schindler, Michael J. Black, and Silvia Zuffi. BITE: Beyond priors for improved three-D dog pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8867–8876, 2023.

Christopher Ruff. Variation in human body size and shape. *Annual Review of Anthropology*, 31: 211–232, 2002. doi: 10.1146/annurev.anthro.31.040402.085407.

Nataniel Ruiz, Miriam Bellver, Timo Bolkart, Ambuj Arora, Ming C. Lin, Javier Romero, and Raja Bala. Human body measurement estimation with adversarial augmentation. In *International Conference on 3D Vision (3DV)*, pp. 219–230, 2022.

Valentin Schmuker, Alex Hoi Hang Chan, Bastian Goldluecke, and Urs Waldmann. Towards texture- and shape-independent 3D keypoint estimation in birds. *arXiv preprint arXiv:2505.16633*, 2025.

Catherine Sheard, Montague H. C. Neate-Clegg, Nico Alioravainen, Samuel E. I. Jones, Claire Vincent, Hannah E. A. MacGregor, Tom P. Bregman, Santiago Claramunt, and Joseph A. Tobias. Ecological drivers of global gradients in avian dispersal inferred from wing morphology. *Nature Communications*, 11:2463, 2020.

Samuel Stevens, Jiaman Wu, Matthew J. Thompson, Elizabeth G. Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M. Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP: A vision foundation model for the tree of life. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19412–19424, 2024.

Chunyi Sun, Ruyi Zha, Weijian Deng, Junlin Han, Dylan Campbell, and Stephen Gould. AnimalLift: Reconstructing animatable 3D animals from a single image by learning canonical shape, texture, and fur maps. *arXiv preprint arXiv:2609.09513*, 2026.

Adrian L. R. Thomas. On the aerodynamics of birds' tails. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 340(1294):361–380, 1993. doi: 10.1098/rstb.1993.0079.

Joseph A. Tobias, Nathalie Seddon, Claire N. Spottiswoode, John D. Pilgrim, Lincoln D. C. Fishpool, and Nigel J. Collar. Quantitative criteria for species delimitation. *Ibis*, 152(4):724–746, 2010. doi: 10.1111/j.1474-919X.2010.01051.x.

Joseph A. Tobias, Catherine Sheard, Alex L. Pigot, et al. AVONET: morphological, ecological and geographical data for all birds. *Ecology Letters*, 25(3):581–597, 2022. doi: 10.1111/ele.13898.

Kuen-Han Tsai and Hsuan-Tien Lin. Learning from label proportions with consistency regularization. In *Asian Conference on Machine Learning (ACML)*, pp. 513–528, 2020.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 595–604, 2015.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8769–8778, 2018.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Urs Waldmann, Alex Hoi Hang Chan, Hemal Naik, Máté Nagy, Iain D. Couzin, Oliver Deussen, Bastian Goldluecke, and Fumihiro Kano. 3D-MuPPET: 3D multi-pigeon pose estimation and tracking. *International Journal of Computer Vision*, 132(10):4235–4252, 2024.

Yufu Wang, Nikos Kolotouros, Kostas Daniilidis, and Marc Badger. Birds of a feather: Capturing avian shape models from images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14734–14744, 2021.

Brian C. Weeks, Zhizhuo Zhou, Bruce K. O'Brien, Rachel Darling, Morgan Dean, Tiffany Dias, Gemmechu Hassena, Mingyu Zhang, and David F. Fouhey. A deep neural network for high-throughput measurement of functional traits on museum skeletal specimens. *Methods in Ecology and Evolution*, 14(2):347–359, 2023. doi: 10.1111/2041-210X.13864.

Brian C. Weeks, Zhizhuo Zhou, Charlotte M. Probst, Jacob S. Berv, Bruce O'Brien, Brett W. Benz, Heather R. Skeen, Mark Ziebell, Louise Bodt, and David F. Fouhey. Skeletal trait measurements for thousands of bird species. *Scientific Data*, 12(1):884, 2025. doi: 10.1038/s41597-025-05234-y.

Heather M. Williams, Samantha B. Wilcox, and Andrea J. Patterson. Photography as a tool for avian morphometric measurements. *Journal of Ornithology*, 161(1):333–339, 2020. doi: 10.1007/s10336-019-01728-w.

Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. DOVE: Learning deformable 3D objects by watching videos. *International Journal of Computer Vision*, 131(10): 2623–2634, 2023a.

Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. MagicPony: Learning articulated 3D animals in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8792–8802, 2023b.

Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, Wei Ji, Chen Wang, Xiaoding Yuan, Prakhar Kaushik, Guofeng Zhang, Jie Liu, Yushan Xie, Yawen Cui, Alan Yuille, and Adam Kortylewski. Animal3D: A comprehensive dataset of 3D animal pose and shape. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9065–9075, 2023.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46: 1212–1230, 2024.

Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. BANMo: Building animatable 3D neural models from many casual videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2853–2863, 2022.

Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. LASSIE: Learning articulated shapes from sparse image ensemble via 3D part discovery. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 15296–15308, 2022.

Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Hi-LASSIE: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4853–4862, 2023.

Anna Zeffer, L. Christoffer Johansson, and Åsa Marmebro. Functional correlation between habitat use and leg morphology in birds (Aves). *Biological Journal of the Linnean Society*, 79(3):461–484, 2003. doi: 10.1046/j.1095-8312.2003.00200.x.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, 2023.

Yivan Zhang, Nontawat Charoenphakdee, Zhenguo Wu, and Masashi Sugiyama. Learning from aggregate observations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7993–8005, 2020.

Silvia Zuffi and Michael J. Black. AWOL: Analysis without synthesis using language. In *European Conference on Computer Vision (ECCV)*, pp. 1–19, 2024.

Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5524–5532, 2017.

Silvia Zuffi, Angjoo Kanazawa, and Michael J. Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3955–3963, 2018.

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

Table 4: **Notation used in the paper**.

Symbol	Meaning
I, $F = \Phi(I)$, f	image, backbone features, pooled feature (mean patch token and class token)
Φ, H, H_0	frozen backbone, regression head, frozen copy of the pretrained head
$\bar{T}$, B, $T(\boldsymbol{\beta})$	AVES template, shape basis, shaped template
$\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$	shape coefficients, bone length scales, joint rotations, camera
W, V, $\hat{V}$, $\mathbf{I}$	linear blend skinning, posed mesh, rest mesh, identity joint rotations of the rest pose
K, Π, x	keypoint regressor, perspective projection, 2D keypoints
ℓ_p, m	length of part p on the mesh, ruler returning the three part-to-wing ratios
$\bar{\ell}^{\,s}$, $\bar{r}^{\,s}$, $\tilde{r}^{\,s}$	mean lengths of species s in AVONET, prior ratios, training targets
k	constant converting the AVONET tail ratio to the tail ruler of the mesh
g_ψ, ψ, c, c_0	proportion head, its parameters, length factors, neutral factors
l, u, u_0, σ	lower and upper bounds of the factors, neutral logit, logistic function
D, o_p, a_p, w_p, ρ_p, s_p	axial deformation, root, axis, part weight, axial profile, gain of part p
$\hat{V}_c$	deformed rest mesh $W(D(T(\boldsymbol{\beta}), c), \boldsymbol{\alpha}, \mathbf{I})$
$\hat{V}^*$	ground-truth rest mesh $\hat{V}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*)$ of a synthetic bird
s, p	species, part (tail, bill or tarsus)
$\mathcal{L}_{\text{2D}}$, $\mathcal{L}_{\text{morph}}$, $\mathcal{L}_{\text{syn}}$	keypoint, morphometric and synthetic ratio losses
$\mathcal{L}_{\text{pose}}$, $\mathcal{L}_{\text{reh}}$	pose and rehearsal losses
$\lambda_m, \lambda_s, \lambda_\theta, \lambda_r$	weights of the loss terms
$\mathcal{B}$, $\mathcal{B}'$	batch of real photographs, batch of synthetic images

A NOTATION

Table 4 lists the symbols used in the paper.

B THE AVONET DATABASE

AVONET (Tobias et al., 2022) is an open database of morphological, ecological and geographical traits for all living birds. It covers 11,009 species in the BirdLife taxonomy, from 243 families in 36 orders, and is also released for the eBird (10,661 species) and BirdTree (9,993 species) taxonomies. We use the BirdLife version throughout.

Raw measurements. The database is built from 90,371 measured individuals: 75,843 museum specimens, mostly from natural history collections, 14,177 live birds measured in the field and 351 of unrecorded type. They were measured by 153 people, 51,354 of them with the AVONET measurement protocol, and 3,309 measurements are taken from published sources. The raw table gives one row per individual with its species, sex, age, source and measurements. The species tables give the average over these individuals.

Traits. Each species has eleven morphological traits. Four are bill measurements: length from the tip to the base of the skull along the culmen, length from the tip to the nostrils, width and depth. The others are tarsus length, wing length, the distance from the carpal joint to the tip of the first secondary feather, Kipp's distance between the tips of the secondary and the longest primary, the hand-wing index derived from it, tail length and body mass. Body mass is taken from the literature and all other traits are measured on the individual. The species tables add ecological and geographical attributes such as habitat, trophic niche, migration and range size, which we do not use.

What we take from it. We use only the four lengths of Table 5. 67,768 individuals of the raw table have all four. For each species we average the four lengths over these individuals and divide tail, bill and tarsus length by wing length (Equation 3). The tail ratio is then multiplied by the constant k of Appendix F. No other trait and no information about the individual in a photograph enters training or evaluation. The variance analysis of Section 3.2 uses the 54,219 individuals of the 6,586 species that have at least five such individuals. The ratios are nevertheless not an identifier. Assigning an individual to its species from the three ratios succeeds 1.75% of the time. Our goal is the morphology of the species, not its recognition.

Table 5: **The four AVONET measurements used in this paper**. Definitions follow Tobias et al. (2022). The last column gives the length on the mesh that transfers each definition (Appendix D).

Trait	AVONET definition	Length on the mesh
Wing length	carpal joint to the tip of the longest primary, on the unflattened wing	ℓ_{wing}, denominator of all ratios
Tail length	tip of the longest tail feather to the point where the two central tail feathers emerge	ℓ_{tail}
Bill length (culmen)	tip of the upper bill to the base of the skull	ℓ_{bill}
Tarsus length	notch of the intertarsal joint to the last scale before the toes	ℓ_{tars}

Matching the image datasets. Each species of CUB and NABirds is matched to one AVONET species. All 200 species of CUB have a match. The 404 species of NABirds are matched by their English name, and 369 of them have exactly one match. We remove the 131 of these that are also CUB species. We also remove Gadwall, a CUB species that our matching had recorded under its former scientific name. This leaves the 237 species of NABirds-237.

C Additional related work

Parametric animal reconstruction. Before single-image regression, SMAL was fitted to images and video (Zuffi et al., 2018; Biggs et al., 2018). Animal3D (Xu et al., 2023) provides a benchmark for mammals that includes fine-tuned human pose and shape estimators. Synthetic training data for animal regressors is also generated in Generative Zoo (Niewiadomski et al., 2025) and Dessie (Li et al., 2024a). Multi-view systems estimate 3D keypoints of birds (Waldmann et al., 2024; Schmuker et al., 2025). AWOL (Zuffi & Black, 2024) maps vision-language embeddings of text or images to the parameters of a quadruped shape space. Surveys are given by Li et al. (2025) and Marvasti-Zadeh et al. (2022).

Template-free reconstruction. CMR (Kanazawa et al., 2018b), MagicPony (Wu et al., 2023b) and 3D Fauna (Li et al., 2024b) learn category shape from image collections without 3D supervision. UCMR (Goel et al., 2020) and DOVE (Wu et al., 2023a) were developed largely on birds. BANMo (Yang et al., 2022), LASSIE (Yao et al., 2022), Hi-LASSIE (Yao et al., 2023) and SAOR (Aygün & Mac Aodha, 2024) learn articulated animals from videos or image collections, and Farm3D (Jakab et al., 2024) learns them from images generated by a 2D diffusion model.

Measurements of bodies and traits. BMnet (Ruiz et al., 2022) estimates anthropometric measurements of humans from silhouettes. Ecomorphology has long related the measurements of birds to their ecology (Ricklefs & Travis, 1980; Leisler & Winkler, 1985), and wing morphology has been linked to dispersal at global scale (Sheard et al., 2020). Across nearly all bird species, nine morphological traits, eight linear measurements and body mass, predict the trophic niche of a species 78% of the time, against 29% for body mass alone (Pigot et al., 2020). Earlier ecomorphological studies point the same way (Miles & Ricklefs, 1984). Avian taxonomy scores differences in culmen, tarsus, wing and tail length as one criterion for species rank (Tobias et al., 2010). Skeletal traits are available for thousands of bird species (Weeks et al., 2025), photography has been evaluated as a tool for avian morphometric measurements (Williams et al., 2020), and whale morphometrics are extracted from drone imagery (Bierlich et al., 2024; Bagchi et al., 2025). In vision, work on biodiversity has centred on recognition (Van Horn et al., 2018; Stevens et al., 2024).

Adaptation and group-level supervision. Adapters (Houlsby et al., 2019) initialize added modules near the identity so that the pretrained network is unaffected when training starts. Learning from a statistic of a group instead of a label for each sample has been studied as learning from aggregate observations (Musicant et al., 2007; Zhang et al., 2020), with learning from label proportions as its counterpart in classification (Quadrianto et al., 2008; Patrini et al., 2014; Tsai & Lin, 2020). In our case the group is a species, the statistic is a mean of measurements taken outside the images, and the target is generalization to groups never seen.

D RULERS AND DEFORMATION DETAILS

This section gives the exact form of the ruler m of Equation 2 and of the operator D of Equation 5. Keypoints are fixed linear combinations of vertices, given by the keypoint regressor K of AVES (Wang et al., 2021). Joints on the rest mesh depend on the bone scales $\boldsymbol{\alpha}$ only. A vertex is said to belong to a joint if its largest skinning weight is on that joint.

Rulers. The wing length is the distance from the wrist joint to the wing-tip keypoint. For the tail, let b be the tail-base keypoint, a the unit vector from b to the tail-tip keypoint, and $\mathcal{T}$ the vertices that belong to a tail joint. The tail length is $\ell_{\text{tail}} = \max_{v \in \mathcal{T}} \max\big(0, \langle v - b, a\rangle\big)$. An earlier ruler used the largest distance from b over $\mathcal{T}$. It also counted vertices on the rump and the width of the fanned tail, and it stopped decreasing when the tail was shortened, so we replaced it. For the bill, consider the direction from the lower-bill joint to the bill-tip keypoint. The root o is the rearmost vertex of the lower bill along this direction, and a is the unit vector from o to the bill-tip keypoint. The bill length is the largest value of $\langle v - o, a\rangle$ over the vertices of the upper and lower bill. For the tarsus, j_1 and j_2 are the two joints that bound the tarsus and a is the unit vector from j_1 to j_2. The length is $\langle y - j_1, a\rangle$ for the foot keypoint y, averaged over the two legs.

Tail. With b, a and $\mathcal{T}$ read from the shaped template, every $v \in \mathcal{T}$ moves to $v + c_{\text{tail}} \max\big(0, \langle v - b, a\rangle\big)\, a$. Vertices behind the tail base and vertices outside $\mathcal{T}$ do not move.

Tarsus. For each leg, with j_1, j_2 and a as above and $L = \|j_2 - j_1\|$, a vertex moves to $v + (c_{\text{tars}} - 1)\, \min\big(L, \max(0, \langle v - j_1, a\rangle)\big)\, w(v)\, a$. The weight $w(v)$ is the sum of the skinning weights of v on the two tarsus joints. The profile saturates at the end of the bone, so the foot is translated as a whole. Both legs use the same factor.

Bill. The field of the bill is computed on the rest mesh. Let $L = \langle x_{\text{tip}} - o, a\rangle$ and let $t(v) = \langle v - o, a\rangle / L$ be the normalized axial position of a vertex. The profile is

$$g(t) = \begin{cases} 0, & t \le t_0, \\ (t - t_0)^2/(2R), & t_0 < t < t_0 + R, \\ t - t_0 - R/2, & t \ge t_0 + R, \end{cases} \tag{9}$$

which is continuous with a continuous derivative. The onset is $t_0 = \max(0.10,\ t_{\text{head}} + 0.05)$, where t_{head} is the largest axial position among the vertices that support the eye, forehead, crown, nape and throat keypoints. These vertices therefore never move. The displacement is $c_{\text{bill}}\, L\, g\big(t(v)\big)\, w(v)\, a$, where $w(v)$ is the sum of the skinning weights on the head and jaw joints, which carry the upper and lower bill, so that both halves of the bill move together. We use $R = 0.35$. The three deformations are applied in the order tarsus, bill, tail, and all displacements are added to the shaped template. Since skinning at the identity pose translates each vertex, they appear unchanged on the rest mesh.

Checks of the bounds. The bounds of Section 3.5 were set before training. We deformed three sets of meshes over the whole range of each factor: the mean bird, 50 meshes predicted for held-out photographs, and one predicted mesh per NABirds species. The criteria were written down before the checks. Body vertices must not move, and the ruler must be monotone in the factor. The 99th percentile of edge stretch at the boundary of the part must be at most 1.5, and the smallest triangle area ratio at least 0.10. Each criterion has a strict version, which every mesh must pass. It also has a main version, which uses the median over images and is the statistic of the mesh gate in Section 4.1. For the lower bound of the tail the strict version allows -0.70 and the main version -0.85. We adopted the latter. At this bound up to 2 triangles of a mesh are flipped. For the bill we adopted the range $[-0.6, 2.0]$ although two strict criteria were not met on some meshes. The tip of the upper bill moves slightly less than the tip of the lower bill, and one or two triangles flip. We record both as known defects of the operator. The statistics of the trained models in Tables 3 and 6 are medians over images.

E IMPLEMENTATION DETAILS

The proportion head takes the 2560-dimensional vector f. It is a layer normalization followed by three fully connected layers with 256, 64 and 3 units and GELU activations (Hendrycks & Gimpel, 2016). The bounds $[l, u]$ are $[-0.85, 4]$ for the tail, $[-0.6, 2]$ for the bill and $[0.15, 4]$ for the tarsus,

with neutral factors 0, 0 and 1. They were fixed before training from checks of the deformed mesh over the whole range. We train with AdamW (Loshchilov & Hutter, 2019) and weight decay 10^{-4}. The learning rate is 10^{-3} for ψ and 3×10^{-5} for H, with linear warm-up over the first 5% of steps and cosine decay.

F THE TAIL RULER CONSTANT k

Why a constant is needed. AVONET measures the tail from the point where the central feathers emerge from the skin to the tip of the longest feather. On the mesh the tail is a surface that starts at the tail-base keypoint, and the two intervals do not coincide. When the tail of the AniMer+ mesh is set to the AVONET ratio itself, it overshoots the tail tip seen in the photograph. On the held-out photographs the PCK of the tail tip is then 3.5, against 14.5 for AniMer+ (An et al., 2026).

Estimate. The pretrained reconstructor is held fixed and nothing is trained. For every training photograph we vary the tail factor on a grid from the lower to the upper bound with step 0.05. We keep the factor whose projected tail tip is closest to the tail tip extracted from the mask. A photograph is used if this distance is within the threshold of PCK@0.1 and the best factor is not at an end of the grid. This leaves 1,958 of 4,075 photographs. The constant is the median, over these photographs, of the tail ratio of the mesh at the best factor divided by the AVONET ratio of the species. This gives $k = 0.445$, with an interquartile range of 0.33 to 0.58 over photographs. Only training species are used.

Criteria fixed before the estimate. We accepted a global constant only if three conditions held. The estimate had to be at most 0.85, that is, clearly different from one. The species-level Pearson correlation between the tail ratio read from the photographs and the AVONET ratio had to be at least 0.5 on the training species. On the held-out species, the PCK of the tail tip with the tail set to k times the AVONET ratio had to be at least that of AniMer+. All three held. The correlation is 0.71 over 140 species, and the PCK of the tail tip is 22.4 against 14.5. The held-out species play no part in the estimate. For reference, the same procedure on the held-out photographs gives a median of 0.431.

What depends on k. Correlations with the measured ratios are invariant to k, and so are all results for bill and tarsus. The proportion error of the tail is measured against k times the AVONET ratio and therefore depends on it. The spread of the estimate over photographs is large, so a single constant is an approximation. Training without it lowers the PCK of the tail tip below that of AniMer+ (Table 6). Bill and tarsus have no counterpart of the tail tip in the photographs, so no such calibration is possible for them, and they are compared with AVONET directly.

G FURTHER RESULTS

Table 8 lists the criteria fixed before each experiment and their outcomes.

Part-specific checks. Two checks are specific to one part, because the general metrics are blind there. The tail tip is not among the CUB keypoints, so PCK does not register the length of the tail. The bill is the only part whose deformation crosses a skinning boundary, where edges can stretch without changing the smallest triangle area. The tarsus needs neither, since the leg keypoints enter PCK and the legs are thin enough for the area ratio to register any distortion.

Selection of the examples. The held-out species of Figure 4 follow a fixed rule. For each part we take the two species with the smallest and the two with the largest AVONET ratio among the held-out species that have at least three test photographs in which bill, tail and at least one leg are visible. Figure 4 shows the first species of the short-tail, long-tail, long-bill and long-tarsus groups. Its photograph is the one of these whose ratios from our model (seed 2) are closest to the species median.

Figure 5 shows five of the 237 NABirds species that no component has seen. For each of tail, bill and tarsus we list the species in the lowest and highest fifth of the AVONET ratio for which our error on that part, averaged over three seeds, is below 10% and that of AniMer+ above 30%. From these lists we chose by hand species that differ in shape and whose other parts also move toward the measurements. For the short tail, the two species with the largest gain are swans, whose bill our model makes far too long, so we show the Laughing Gull instead, whose tail error is 12%. Long

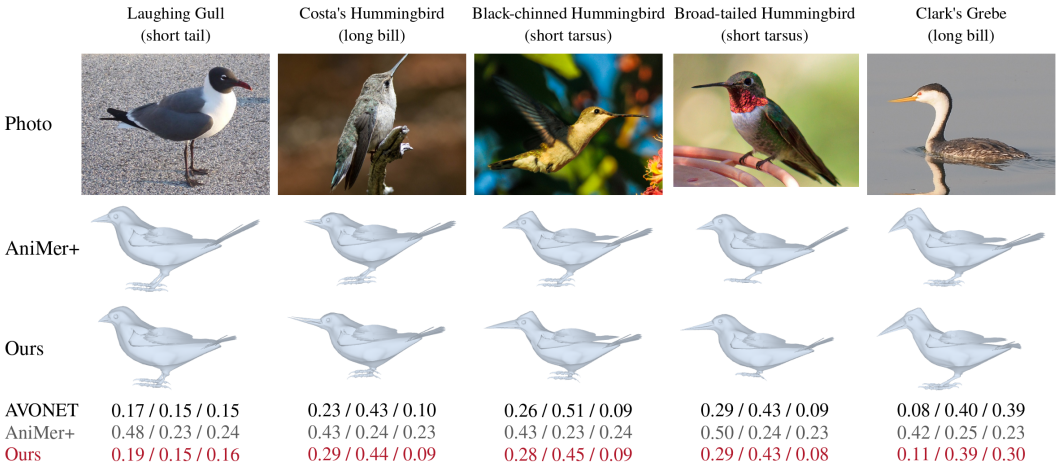


Figure 5: **Reconstructions of five species unseen in training**. NABirds photographs, the rest meshes of AniMer+ and of our model (seed 0) at equal wing length, and tail / bill / tarsus length relative to wing length measured in AVONET and on each mesh. For each species the photograph is the one whose ratios are closest to the species median of our model, except for the gull, whose closest photograph carries a watermark and is replaced by the second closest. Photographs by Sam Galick, Aaron, Shelly Spencer, Muriel Neddermeyer and Cory Gregory (left to right).

Table 6: **All values of the alternatives and the component ablations**. Upper rows: the alternatives of Table 3. Lower rows: each removes or relaxes one component of the full model. Error: proportion error (%). PCK: PCK@0.1 on CUB. Tip: PCK of the tail tip on CUB-49. PA-V: PA-MPVPE on synthetic birds. Area: smallest triangle area ratio. Seam: edge stretch at the boundary of the deformed bill. A value marked $\times$ fails the check of its column relative to our model (Section 4.3), and Fails lists the failed checks. Tip is marked where it falls below that of AniMer+ (14.5). n: number of seeds. The last row is trained with the tail target without the constant k, and its tail error is measured against the k-scaled target like every other row.

		Error (%) ↓		CUB ↑		Synthetic ↓	Mesh		
Method	n	CUB-49	NAB-237	PCK	Tip	PA-V	Area ↑	Seam ↓	Fails
Ours	3	10.3	14.0	0.643	30.2	707.5	0.64	1.00	none
Ratio loss on $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, anchor.	3	24.9$^\times$	36.4$^\times$	0.632	21.1	704.0	0.38	1.10	error
Ratio loss on $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, free	3	18.0$^\times$	30.5$^\times$	0.667	21.8	719.8	0.04$^\times$	1.61$^\times$	error, mesh
2D keypoints only	3	119.1$^\times$	127.3$^\times$	0.655	37.9	706.1	0.40	2.60$^\times$	error, mesh
Stretching after recon.	3	9.1	14.8	0.594$^\times$	20.0	827.5$^\times$	0.59	1.00	PCK, 3D
w/o $\mathcal{L}_{\text{syn}}$	3	10.2	14.1	0.642	30.5	804.4$^\times$	0.64	1.00	3D
Shape anchored, not frozen	3	10.7	14.1	0.638	30.4	699.3	0.38	1.04	none
Shape free	3	10.4	14.0	0.674	29.5	699.0	0.05$^\times$	1.77$^\times$	mesh
w/o tail constant k	1	50.8$^\times$	54.1$^\times$	0.622	10.9$^\times$	706.9	0.66	1.00	error

tails are not shown because the default tail of AniMer+ is already close to them. AniMer+ gives the five birds nearly the same proportions. Our model shortens the tail, the bill and the legs of the gull, lengthens the bill and shortens the legs of the hummingbirds, and shortens the tail and lengthens the bill and the legs of the grebe.

Components. Table 6 shows what each component contributes. Without $\mathcal{L}_{\text{syn}}$ the proportion error stays that of our model, but PA-MPVPE rises from 707.5 to 804.4, because the head deforms synthetic birds toward species averages. Letting body shape move does not improve the proportions. Anchored, it passes every check but lowers the area ratio to 0.38. Free, the mesh leaves the gate although PCK and PA-MPVPE improve. We therefore keep body shape frozen. Without k the tail target is too long for the mesh, the tail error against the k-scaled target rises to 128%, and the PCK of the tail tip drops to 10.9, below AniMer+. Table 7 gives the values for each part and the standard deviations over seeds.

Table 7: **Ablations with values for each part and standard deviations over seeds**. Proportion error on CUB-49 for tail, bill and tarsus (%), PCK@0.1, PCK of the tail tip and PA-MPVPE, as mean and standard deviation over n seeds. Rows A are the alternatives and rows B the component ablations of Table 6, followed by three further settings of our model that were run with one seed.

	Method	n	Tail ↓	Bill ↓	Tars. ↓	PCK ↑	Tip ↑	PA-V ↓
	AniMer+	1	31.3	21.0	27.9	0.594	14.5	709.0
	Ours	3	8.9 ±0.6	11.1 ±1.1	11.0 ±1.1	0.643 ±0.001	30.2 ±1.3	707.5 ±0.8
A	Ratio loss on $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, anchored	3	27.1 ±0.1	20.1 ±0.4	27.4 ±0.1	0.632 ±0.000	21.1 ±0.6	704.0 ±2.1
	Ratio loss on $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, free	3	23.3 ±0.2	17.3 ±0.2	13.4 ±0.4	0.667 ±0.002	21.8 ±0.4	719.8 ±1.7
	2D keypoints only	3	21.5 ±0.9	178.2 ±0.2	157.6 ±9.6	0.655 ±0.001	37.9 ±0.3	706.1 ±1.0
	Stretching after reconstruction	3	7.7 ±0.8	9.4 ±1.1	10.3 ±1.5	0.594 ±0.000	20.0 ±0.6	827.5 ±4.0
B	w/o $\mathcal{L}_{\text{syn}}$	3	8.7 ±1.1	11.1 ±0.8	10.8 ±0.2	0.642 ±0.000	30.5 ±0.2	804.4 ±3.6
	Shape anchored, not frozen	3	9.4 ±1.0	11.8 ±1.2	11.0 ±0.3	0.638 ±0.001	30.4 ±1.0	699.3 ±0.4
	Shape free	3	8.8 ±1.0	11.1 ±1.3	11.2 ±0.2	0.674 ±0.001	29.5 ±0.7	699.0 ±0.6
	w/o tail constant k	1	128.3	12.6	11.4	0.622	10.9	706.9
	Shape weakly anchored	1	7.9	12.5	10.9	0.646	30.7	699.2
	3,000 training steps	1	8.4	10.4	12.3	0.636	29.9	704.5
	w/o tail-base keypoint	1	9.5	9.8	12.1	0.633	32.3	708.6

Comparison with WildProp. Six species are shared with the evaluation of WildProp (Chasmai et al., 2026), and we compare the tarsus proportion on them. Its mean error on them is 10.5%, ours is 19.5% and that of AniMer+ is 36.5%. The gap comes from the two long-legged species, Great Egret and Lesser Yellowlegs (for the egret, 49.9% against 8.6%). On three of the four other species our error is the lower one (Table 9). For these four shorebirds WildProp is evaluated against another dataset, so the comparison is indicative. WildProp uses an annotated query image and a collection of photographs of the species, while we reconstruct a mesh from one photograph.

Stretching after reconstruction. The baseline of Table 3 uses a small network on the same frozen features that predicts the three ratios. It is trained on the species averages of the training species. The number of epochs is chosen on 30 training species set aside for this purpose, and the network is then retrained on all training species, so that no held-out or unseen species enters its training or selection. We train it with three seeds and report the mean. The factors of the operator of Appendix D are then solved so that the ruler on the AniMer+ mesh reads the predicted ratios, with the tail ratio multiplied by k. The reconstructor is not trained.

Training length. The check of Section 4.1 was run on an earlier variant of our model without $\mathcal{L}_{\text{syn}}$. After 1,500, 3,000 and 4,500 steps the mean proportion error is 9.52, 6.66 and 6.26% on the validation set and 11.68, 11.32 and 10.50% on CUB-49. The validation species are training species, so this set measures generalization to new photographs and not to new species.

First round of adding species. Section 4.5 reports the second round of this experiment. A first round, with an earlier variant of our model and one seed, gave errors of 12.71, 12.73, 11.54 and 11.10% on the never-shown species. The error did not decrease at the first step, so its criterion was not met, and we repeated the experiment with the final model and three seeds. An earlier experiment that added only six heron species and eight short-legged species lowered the tarsus error of twelve long-legged species from 40.1 to 19.0%. It also lowered the correlations on CUB-49 from 0.88, 0.80 and 0.90 to 0.83, 0.77 and 0.87. Adding species across families, as in Section 4.5, avoids this cost.

What breaks the mesh. To see why the pathway lies outside the shape space, we exchange parameters between models on the 678 held-out photographs and measure the rest mesh. The shape route of Table 3 (free variant) is the model whose $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ were moved by the ratio loss. Starting from the parameters predicted by AniMer+, we replace $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ or both by those of the shape route, or add the factors of our pathway. On each photograph we check the three conditions of the mesh gate (smallest area ratio at least 0.10, edge stretch at most 1.5, no negative bone scale) and whether any triangle is flipped. Table 10 gives the result over three seeds. Moving $\boldsymbol{\beta}$ or $\boldsymbol{\alpha}$ toward the measured proportions breaks the mesh on 97 to 100% of the photographs. The norm of $\boldsymbol{\beta}$ rises from 1.0 to 3.1, 98% of the photographs have a negative bone scale when $\boldsymbol{\alpha}$ is moved, and the proportion error is still 18.0%. The pathway leaves $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ untouched and reaches 10.3% with a median area ratio of 0.64. It is not free of defects. We had fixed a criterion of fewer than 5% failing photographs for the

Table 8: **Criteria fixed before the runs and their outcomes**. Each criterion was written down before the corresponding experiment was run and was not changed afterwards.

Experiment	Criterion	Outcome
Mesh gate	Smallest area ratio at least 0.10, edge stretch at most 1.5, no negative bone scale, as medians over images	Applied to all models
Tail constant k	The three conditions of Appendix F	Met
Ratio loss on $(\boldsymbol{\beta}, \boldsymbol{\alpha})$	The route is called worse than the pathway if, for every seed, the mean error exceeds ours by 2.0 points, or the mesh gate fails, or PCK@0.1 falls by 0.02 or PA-MPVPE rises by 5%	Met for both settings and all seeds
2D keypoints only	For every seed, r of bill and tarsus below 0.3 and edge stretch of the bill above 1.5	Met
Shape free	For every seed, r within 0.03 of ours and the mesh gate failed	Met
Training length	The validation set alone prefers 4,500 to 3,000 steps, and validation and held-out species rank the two in the same order	Met (earlier variant)
Adding species	Error on the never-shown species decreases with each step, and with three seeds the gain from none to all exceeds the sum of the two standard deviations. CUB-49 error rises by at most 1.0 point. Mesh gate passed	Not met with the earlier variant (one seed); met with the final model (three seeds)
Weight for extreme species	Bill and tarsus slopes on NABirds rise by at least 0.15	Not met
Target range	At least one correlation on NABirds rises by two seed standard deviations	Not met
Morphospace	Hull area of our species relative to the measured one at least three times that of AniMer+	Met
2D keypoints on NABirds	PCK@0.1 of every seed at least that of AniMer+	Met
Signed-rank test	$p < 0.01$ for every part, seed and test set	Met
Orders	Our error lower in at least 80% of the orders with at least three species	Met (13 of 13)
Families absent from training	Error below that of AniMer+, and r at least 0.5 for at least two parts	Met
Lookup by family	Our r at least that of the lookup minus 0.02 for at least two parts	Met (tail, tarsus)
Consistency within a species	Share among species above that of AniMer+ and at least 0.5	Met
Within natural variation	Our count at least twice that of AniMer+ for every part	Not met (NABirds bill, 1.9 times)
CowBird	Our error below that of AniMer+ for every seed in at least two parts	Met (tail, bill)
Individual ratios of synthetic birds	Our r above that of AniMer+ by 0.1 for at least two parts	Not met

pathway, and it is not met. On 29% of the photographs, mostly those in which the tail is shortened strongly, a few triangles inside the fan of the tail flip: a median of 6 and at most 79 of the 14,268 faces. The per-photograph failures of the gate conditions (20%) come from the area condition on the same strongly shortened tails. The gate itself, applied to medians, is passed. This is the defect noted in Appendix D.

Two remedies for the compression that did not help. Section 4.4 shows that extreme proportions are compressed on unseen species. We tried two changes of the loss. Both were run on an earlier variant of our model, without $\mathcal{L}_{\text{syn}}$ and with body shape anchored, and their criteria were fixed before the runs. The first gives a larger weight to species with extreme targets. The weight of a photograph is $1 + |z|$, capped at 4, where z is the standardized log target averaged over the three parts. The criterion was a rise of the bill and tarsus slopes on NABirds by at least 0.15. With the weight the slopes are 0.71, 0.40 and 0.45 for tail, bill and tarsus, against 0.69, 0.43 and 0.44 without it, so the criterion is not met. The second replaces the target mean by a target range. The loss is zero while the log ratio lies within one within-species standard deviation of the species mean. The criterion was a rise of at least one correlation on NABirds by two seed standard deviations. The

Table 9: **Tarsus proportion error on the six species shared with WildProp**. Relative error of the tarsus-to-wing ratio (%). Ours is measured against AVONET. WildProp values are those reported by Chasmai et al. (2026), against AVONET for Great Egret and White-tailed Kite and against the shorebird measurements of the Alaska Science Center for the other four species. Ours is the mean of three seeds.

Species	WildProp ↓	Ours ↓
Great Egret	8.6	49.9
Lesser Yellowlegs	5.4	30.3
Western Sandpiper	15.4	20.1
Dunlin	12.8	11.8
Ruddy Turnstone	5.2	2.5
White-tailed Kite	15.7	2.4
Mean	10.5	19.5

Table 10: **Mesh integrity when parameters are exchanged between models**. Rest meshes of the 678 held-out photographs. Fail: photographs on which at least one condition of the mesh gate fails (%); the gate itself is applied to medians over photographs (Section 4.1). Area and stretch: medians over photographs. Flip and negative: photographs with at least one flipped triangle or a negative bone scale (%). Err.: mean proportion error on CUB-49 (%). $\|\boldsymbol{\beta}\|$: median norm of the shape coefficients. Values for the last four rows are means over three seeds.

Parameters	Fail ↓	Area ↑	Stretch ↓	Flip ↓	Negative ↓	Err. ↓	$\|\boldsymbol{\beta}\|$
AniMer+	0.0	1.00	1.00	0.0	0.0	26.7	1.0
Pathway only (ours)	20.2	0.64	1.13	29.1	0.0	10.3	1.0
$\boldsymbol{\beta}$ of the shape route	96.7	0.04	1.62	100.0	0.0	22.1	3.1
$\boldsymbol{\alpha}$ of the shape route	100.0	0.08	3.57	100.0	98.2	22.4	3.1
$\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ of the shape route	99.9	0.04	3.63	100.0	98.2	18.0	3.1

correlations are 0.89, 0.71 and 0.63 with the range and 0.89, 0.71 and 0.63 without it. The range is narrower than the error of the prediction, so it acts like a point target. Neither change moves the most compressed species, the long-tailed ones under the weight and the long-legged ones under the range. We conclude that the compression comes from the coverage of the training species and not from the form of the loss, which agrees with Section 4.5.

Caveats of the evaluation. Bill and tarsus are compared with AVONET without any correction, which assumes that the ruler on the mesh matches the measurement taken on a bird in the hand. 3D error can be measured only on synthetic birds, whose proportions do not follow those of real species, so there is no direct 3D check of proportions on real birds. A comparison with species averages cannot tell whether the proportions of one individual are right. Within a species the ratios vary by 6.0–6.5%, which includes differences between sexes and ages. For the silhouette, precision falls while recall rises, that is, the lengthened parts cover area outside the mask. The tail of the template is a wide fan, and the operator has no width term. Adding one is left for future work.

H Further analyses

The analyses of this section use the models of the main paper. Their criteria were fixed before the runs and are listed with their outcomes in Table 8.

Morphospace. Figure 6 places the 237 NABirds species in two planes of the ratios. The first plane is also shown in Figure 3 (bottom). We measure the spread of each set of species by the area of its convex hull, relative to that of the measured species. AniMer+ covers 1.3% of it in the tail–tarsus plane and 0.4% in the bill–tarsus plane, and our model 32.6% and 41.3%. On CUB-49 the values are 1.0% and 0.3% for AniMer+ against 34.0% and 47.9% for ours.

Within natural variation. AVONET also gives the spread of each ratio over the individuals of a species. We count the species whose predicted ratio lies within one standard deviation of the measured mean. Of the 180 NABirds species with at least five measured individuals, these are 66, 50 and 53 for our model (tail, bill, tarsus) and 19, 26 and 20 for AniMer+. For the 40 such species

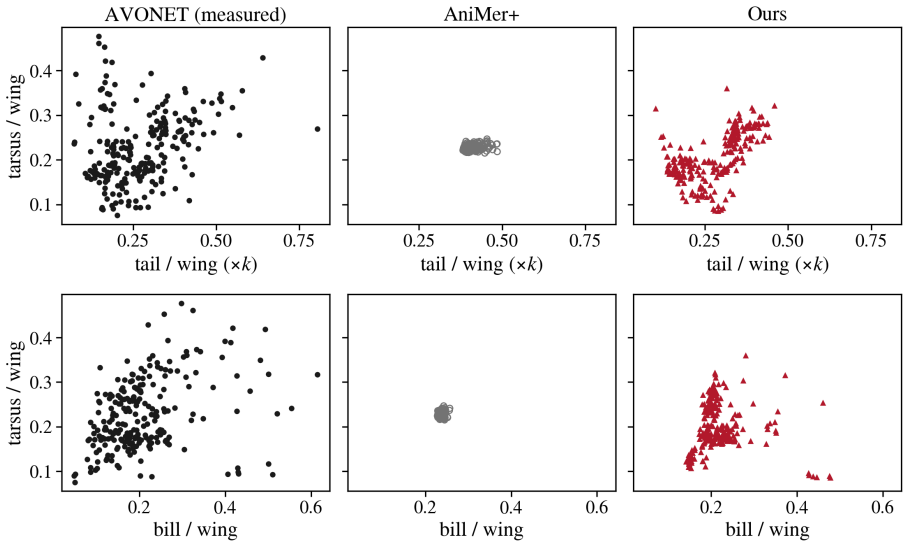


Figure 6: **Morphospace of the 237 NABirds species**. One marker per species: tarsus ratio against tail ratio (top) and against bill ratio (bottom), measured in AVONET (left, black dots), reconstructed by AniMer+ (middle, grey circles) and by our model (right, red triangles, mean of three seeds). All ratios are relative to wing length, and the tail ratio is multiplied by k. Axes are shared within a row.

Table 11: **Proportion error by order on NABirds-237**. Median over the species of an order of the error averaged over tail, bill and tarsus (%), for the orders with at least three species. n: number of species. Ours is the mean of three seeds.

Order	n	AniMer+ ↓	Ours ↓	Order	n	AniMer+ ↓	Ours ↓
Passeriformes	93	22.8	12.0	Galliformes	7	40.7	30.0
Anseriformes	32	85.2	20.0	Strigiformes	7	63.2	42.6
Charadriiformes	32	61.1	16.7	Falconiformes	5	79.7	30.8
Accipitriformes	12	80.7	36.2	Columbiformes	5	46.7	26.6
Pelecaniformes	12	73.3	34.1	Suliformes	4	38.8	9.7
Caprimulgiformes	9	83.7	11.9	Gruiformes	3	66.8	19.8
Piciformes	9	29.3	8.1				

of CUB-49 the counts are 19, 15 and 10 against 2, 7 and 3. The median standard deviation is 4.8 to 7.5% of the mean. With a tolerance of 20%, our ratio is within range for 175, 125, 161 of the 237 NABirds species, against 57, 77, 87 for AniMer+.

Consistency within a species. Section 3.2 reports that 93 to 95% of the variance of the measured log ratios lies among species. The same decomposition of the predicted log ratios over the photographs of NABirds-237 gives a share among species of 0.77, 0.62 and 0.69 for our model and 0.17, 0.36 and 0.34 for AniMer+. On CUB-49 the shares are 0.85, 0.78 and 0.87 against 0.25, 0.33 and 0.36. The predictions vary with the species more than with the photograph.

Ratios without the wing. All three ratios share wing length as denominator. The ratios between the parts themselves, tail to bill, tail to tarsus and bill to tarsus, are also recovered. Their correlations with the measured species ratios on NABirds-237 are 0.79, 0.69 and 0.86 for our model and -0.05, -0.01 and 0.20 for AniMer+, and on CUB-49 0.82, 0.60 and 0.84 against -0.23, 0.05 and 0.10.

Orders. We group the NABirds species by order and compare the median over species of the error averaged over the three parts. Our error is lower in all 13 orders with at least three species (Table 11), and also in all 5 orders with fewer species.

Families absent from training. 115 of the 237 NABirds species belong to one of 31 families that have no species among the training species. On them the error falls from 76.4, 46.8 and 22.8% to 12.9, 32.9 and 18.7%, and the correlation rises from -0.42, 0.72 and 0.40 to 0.85, 0.81 and 0.44. The tarsus correlation stays low. Herons and sandpipers are among these species.

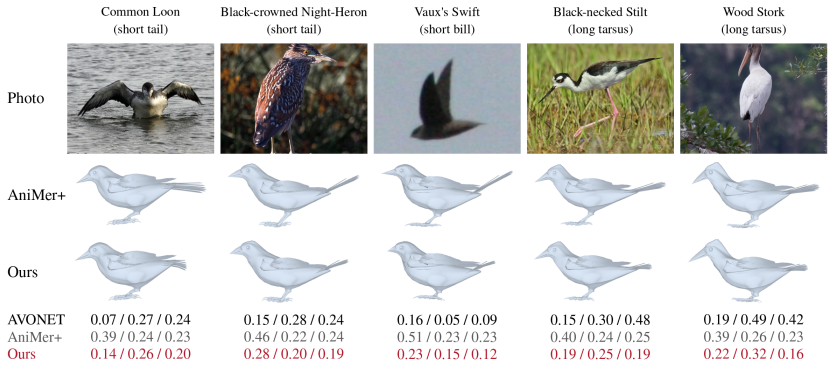


Figure 7: **Reconstructions of the NABirds species with the largest error**. Species of NABirds-237 with a large error of our model (mean of three seeds): the two largest for the tail, the largest for the bill, and two long-legged species among the four largest for the tarsus. Photograph, meshes and ratios as in Figure 5. Photographs by Velma Knowles, Michael L. Baird, Andrew Core, Ryan Schain and Jon Corcoran (left to right).

Lookup by family. A stronger baseline than that of Section 4.4 knows the species name. For each NABirds species with a family among the training species, it returns the mean AVONET ratio of the training species of that family. On these 122 species its correlations are 0.90, 0.79 and 0.91 and ours 0.93, 0.72 and 0.90, with errors of 12.0, 11.4 and 8.7% against 10.5, 10.9 and 9.0%. From one photograph and without the species name, our model is better for the tail, on par for the tarsus, and for the bill has a lower correlation but a lower error.

2D keypoints on NABirds. NABirds annotates eight points that have a counterpart among the keypoints of the model: bill, crown, nape, both eyes, back, breast and tail. On the 4,740 photographs of NABirds-237, with the same crop as in our evaluation, PCK@0.1 rises from 0.761 to 0.783 and PCK@0.15 from 0.884 to 0.901, and every seed is above AniMer+. The largest gains are on the bill (0.522 to 0.574) and the tail (0.397 to 0.457).

Significance of the error reduction. For the error, a Wilcoxon signed-rank test over species of the difference between the error of AniMer+ and ours gives $p \leq 2.9 \times 10^{-6}$ on CUB-49 and $p \leq 1.1 \times 10^{-12}$ on NABirds-237, for every part and seed.

A dataset outside training. The CowBird test set (Badger et al., 2020) contains 114 instances of one species, the Brown-headed Cowbird, which is not a CUB species. Against its AVONET ratios, the median error over photographs falls from 34.8% to between 11.9 and 13.4% for the tail over the three seeds and from 38.3% to between 12.3 and 21.7% for the bill. For the tarsus it rises from 6.5% to between 18.4 and 23.4%. The default tarsus of AniMer+ happens to be close to that of this species. With the evaluation of AniMer+ on this set (An et al., 2026), PCK@0.15 rises from 0.284 to 0.309 and the AUC from 0.603 to 0.610.

Individual ratios of synthetic birds. The synthetic test images have ground-truth meshes. The correlation between predicted and true ratios over the 309 images is 0.87, 0.46 and 0.72 for our model and 0.87, 0.44 and 0.72 for AniMer+. Without $\mathcal{L}_{\text{syn}}$ it drops to 0.64, 0.09 and 0.04, with errors of 24.1, 20.2 and 18.1%. The term keeps the individual proportions that the species prior would overwrite.

Failure cases. Figure 7 shows NABirds species on which our error is large. For tail and bill they are the species with the largest error. For the tarsus we show two long-legged species among the four largest. The largest tarsus error, of the Monk Parakeet, comes from a short tarsus that our model makes too long. Tail and bill change in the right direction but not enough: the tail of the Common Loon goes from 0.39 to 0.14 (measured 0.07), and the bill of Vaux's Swift from 0.23 to 0.15 (measured 0.05). The tarsus of the long-legged Black-necked Stilt and Wood Stork is shortened instead of lengthened (0.25 to 0.19 and 0.23 to 0.16, measured 0.48 and 0.42), as in Section 4.4.