MagicPony: Learning Articulated 3D Animals in the Wild
– Supplementary Material –

Shangzhe Wu* Ruining Li* Tomas Jakab* Christian Rupprecht Andrea Vedaldi
Visual Geometry Group, University of Oxford
{szwu, ruining, tomj, chrisr, vedaldi}@robots.ox.ac.uk
3dmagicpony.github.io

1. Additional Results

1.1. Additional Comparisons with Prior Works

Qualitative Comparisons Fig. 5 and Fig. 6 show qualitative comparisons with prior works on both horses and birds. Our method predicts 3D shapes with finer details and more accurate poses, compared to prior works. We also plot the distribution of predicted viewpoints demonstrating that other methods with a comparable level of supervision collapse to only a limited range of viewpoints, *e.g*. predicting only frontal poses.

We also compare against another recent method, LASSIE [8], which also leverages DINO-ViT [1] image features but only *optimises* over a small set of images ($\sim$ 30). As illustrated in Fig. 1, LASSIE [8] starts from a heavily hand-crafted part-based initial shape, whereas our method starts with a generic ellipsoid with only a simple description of the bone topology. Yet after training, our model produces more detailed 3D shapes from a new test image *unseen* at training, compared to the reconstructions obtained by LASSIE which are directly *optimised* on these images.

Visualisations of Toy Bird Reconstructions. Supplementary to Tab. 2 in the main paper, we show a qualitative comparison of the predicted shapes and the scanned ground-truth shapes on the Toy Bird Scan dataset in Fig. 4.

1.2. Ablation Studies

We further provide quantitative ablation studies on the Toy Bird Scans benchmark in Tab. 1, validating the effectiveness of individual components of the model. In addition to Fig. 5 in the main paper, we also examine the effects of both the feature rendering loss $\mathcal{L}_{\text{feat}}$ and the multi-hypothesis viewpoint prediction in Fig. 3, demonstrating that both of the components are essential to prevent the collapse of viewpoint prediction.

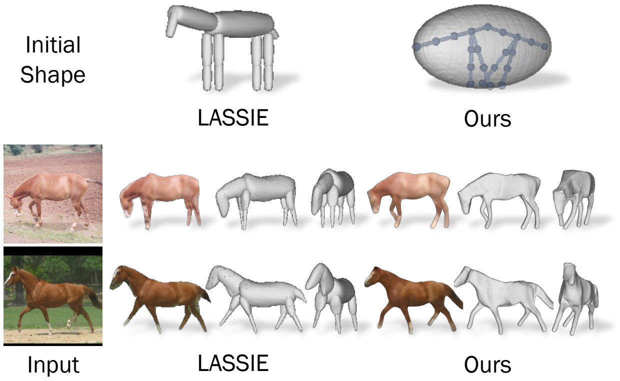

Figure 1. **Comparison with LASSIE [8] on Horses.** LASSIE starts from a heavily hand-crafted part-based initial shape, whereas our method simply starts with a generic ellipsoid with a simple heuristic description of the bone topology. After training, our method produces more plausible 3D shapes from an *unseen* test image, compared to the reconstructions obtained from LASSIE which are directly *optimised* on these images. Note that LASSIE represents the 3D shape with a set of disjoint primitive parts, resulting in unnatural junctions.

1.3. Texture Finetuning

Fig. 7 shows how a quick test-time finetuning (100 iterations) of the predicted texture improves their quality. This is especially effective for images that are far from the training set distribution. Note that the textures of the real horses in the main paper are predictions from a single forward pass.

1.4. Additional Qualitative Results

Additional results of single image reconstruction, animation and relighting can be found in Fig. 8 and the supplementary video. More generalisation results on abstract horse drawings, sculptures and toys are presented in Fig. 9, showing that the model has learned to estimate shape, pose and articulation sufficiently robustly to generalise beyond the training distribution.

[*] Equal contribution.

We also show more reconstruction results of giraffes, zebras and cows in Fig. 10. Our method is able to produce accurate shape reconstructions from a single image across large variations of animal shapes.

1.5. Failure Cases

Our texture prediction might not generalise well enough beyond the distribution of textures observed during training. This is particularly apparent when the trained model is applied on paintings and abstract drawings of horses, neither of which are part of the training set. We demonstrate that a quick finetuning step of the albedo network (100 iterations which takes less than 10 seconds) can remedy this shortcoming. Fig. 7 illustrates the difference between the single-pass predicted textures and the finetuned version.

The viewpoint prediction can fail in the case of more extreme and ambiguous views as shown Fig. 2. This is often caused by DINO-ViT features that are less distinctive for these particular views.

When the horse is observed from a side-view, the method might not be able to disambiguate between left and right legs, for instance, in the second to last row of Fig. 8. Note that our method uses only object masks and self-supervised DINO-ViT features, neither of which are sufficient to disambiguate between different legs of an animal.

2. Additional Technical Details

2.1. Articulation Model Details

Recall that our model estimates a set of bones and articulates the instance mesh using a linear blend skinning model with predicted bone rotations. In the following, we describe in detail how the rest-pose bones are estimated and how the skinning weights are defined.

Bone Topology. Our method only assumes a description of the topology of the animal's skeleton, and automatically estimates a set of bones at rest pose for the articulation model based on simple heuristics. Specifically, for birds, we estimate a chain of 8 bones with equal lengths that lie on two line segments going from the centre (root) of the rest-pose mesh to the two most extreme vertices along z-axis (4 bones on each side), forming a 'spine'.

For quadrupedal animals, like horses, we lift the root joint slightly higher and further add 4 sets of bones for modelling the legs, as illustrated in Fig. 1. We first identify the foot joints as the lowest points of mesh (in y-axis) in each of four xz-quadrants. We then draw 4 line segments from the foot joints to their closest spine joints, and define a chain of 3 bones with equal lengths on each of the segments, representing each leg.

Skinning Weights. Recall Eq. (2) in the main paper, where the instance mesh is further posed by a linear blend skinning equation. Each vertex $V_{\text{ins},i}$ is associated with the bones by a skinning weight $w_{i,b}$, defined as:

$$\begin{aligned} w_{i,b} &= \frac{e^{-d_{i,b}/\tau_\text{s}}}{\sum_{k=1}^{B} e^{-d_{i,k}/\tau_\text{s}}}, \\ \text{where} \quad d_{i,b} &= \min_{r\in[0,1]} \|V_{\text{ins},i} - r\tilde{\mathbf{J}}_b - (1-r)\tilde{\mathbf{J}}_{\pi(b)}\|_2^2 \end{aligned} \tag{1}$$

is the minimal distance from the vertex $V_{\text{ins},i}$ to each bone b, defined by the rest-pose joint locations $\tilde{\mathbf{J}}_b, \tilde{\mathbf{J}}_{\pi(b)}$ in world coordinates, and τ_s is a temperature parameter set to 0.5.

Table 1. Ablations on Toy Bird Scans (Chamfer Distance ↓ in cm). Annotations: 🕘 τ schedule, 🦴 articulation, ◑ symmetry.

full	w/o $\mathcal{R}_\text{Eik}$	w/o $\mathcal{R}_\text{def}$	w/o 🕘	w/o ◑	w/o 🦴 ◑	w/o $\mathcal{L}_\text{im}$	4 / 16 bones
0.79	1.19	1.37	1.66	1.56	1.63	2.11	2.10 / 1.27

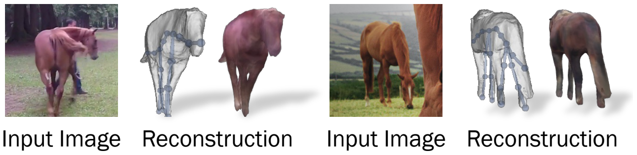

Figure 2. **Incorrect Viewpoint Predictions.** The viewpoint prediction can be less reliable in the case of more extreme input views.

Constraints on the Bone Rotations. Our model learns complex articulated 3D poses of animals using reconstruction losses on single-view images, without any explicit 3D geometric supervision, which is an extremely ill-posed task. In order to prevent unnatural poses, we enforce minimal constraints on the bone rotations: (1) all bone rotations are limited to $(-60°, 60°)$, (2) for quadrupeds, leg rotations around y- and z-axes ('twist' and 'side-bending') are further limited to $(-18°, 18°)$.

2.2. Network Architectures

We implement the feature field ψ, template SDF s and the light network f_l using 5-layer MLPs, and the albedo field f_a and deformation field $f_{\Delta V}$ with 8-layer MLPs. The articulation network consists of 4 transformer blocks. All coordinate inputs are encoded using $\texttt{sin}(\cdot)$ and $\texttt{cos}(\cdot)$ with 8 frequencies.

The encoders are simple convolutional networks, described in Sec. 2.4. In practice, the viewpoint network f_vp is also (separately) implemented using the same architecture as the encoder. Abbreviations of the components are defined as follows:

- Conv$(c_{in}, c_{out}, k, s, p)$: 2D convolution with c_{in} input channels, c_{out} output channels, kernel size k, stride s and padding p
- GN(n): group normalization [7] with n groups

Table 2. Architecture of the patch feature encoders f_k, f_o.

Encoder	Output size
Conv(384, 256, 4, 2, 1) + GN(64) + LReLU(0.2)	16 × 16
Conv(256, 256, 4, 2, 1) + GN(64) + LReLU(0.2)	8 × 8
Conv(256, 256, 4, 2, 1) + GN(64) + LReLU(0.2)	4 × 4
Conv(256, 256, 4, 2, 0) → output	1 × 1

- LReLU(p): leaky ReLU [5] with a slope p

2.3. Hyper-parameters and Training Details

All hyper-parameters are listed in Tab. 3. We enable the articulation after 10k iterations and the deformation after 40k iterations, to prevent the model from overfitting individual images with excessive articulation and deformation. During the first 5k iterations, we allow the model to explore all four viewpoint hypotheses by randomly sampling the four hypotheses uniformly, and gradually decrease the chance of random sampling to 20% while sampling the best hypothesis for the rest 80% of the time. The temperature τ is decreased from 1 to 0.01 over the course of 100k iterations. It takes roughly 20 hours to train the full model for 150k iterations on one single NVIDIA A40 GPU.

2.4. Keypoint Transfer Evaluation Details

Due to the lack of 3D ground-truth for in-the-wild objects, we employ the Keypoint Transfer task and compute the Percentage of Correct Keypoints (PCK) [2–4] as an indirect metric for evaluating the reconstructed 3D shapes, as described in Sec. 4.4 in the main paper. A transferred keypoint is correct if it is within a distance d of the corresponding ground-truth 2D keypoint in the target image. The value of d is computed as $0.1 \cdot \max(h, w)$ for PCK@0.1, where h and w represent the height and width of the ground-truth object bounding box. We follow the open-source implementation[1] of the metric as described in [3] for the PASCAL VOC Horse dataset and in [4] for the CUB Bird dataset. It should be noted that [3] defines their error with respect to a bounding box that is padded by 5% of the original size on each side. To maintain consistency, we follow the same practice for the PASCAL VOC Horse dataset.

Table 3. Training details and hyper-parameter settings.

Parameter	Value/Range
Optimiser	Adam
Learning rate on prior (ψ and s)	1×10^{-3}
Learning rate on others	1×10^{-4}
Number of iterations	150k
Batch size	10
Loss weight λ_m	10
Loss weight λ_dt	10
Loss weight λ_im	1
Loss weight λ_f	10
Loss weight λ_E	0.01
Loss weight λ_d	10
Loss weight λ_a	0.1
Loss weight λ_h	1
Image size	256 × 256
Field of view (FOV)	25°
Camera location	$(0, 0, 10)$
Tetrahedral grid size	256
Initial mesh centre	$(0, 0, 0)$
Translation in x- and y-axes	$(-0.4, 0.4)$
Translation in z-axis	$(-1.0, 1.0)$
Number of spine bones	8
Number of bones for each leg	3
Viewpoint hypothesis temperature τ	$(0.01, 1.0)$
Skinning weight temperature τ_s	0.5
Ambient light intensity k_s	$(0.0, 1.0)$
Diffuse light intensity k_d	$(0.5, 1.0)$

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1

[2] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 3

[3] Nilesh Kulkarni, Abhinav Gupta, David F. Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *CVPR*, 2020. 3

[4] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *ECCV*, 2020. 3

[5] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 3

[6] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. DOVE: Learning deformable 3d objects by watching videos. *arXiv preprint arXiv:2107.10844*, 2021. 4, 5

[7] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 2

[8] Chun-Han Yao, Wei-Chih Hung, Michael Rubinstein, Yuanzhen Lee, Varun Jampani, and Ming-Hsuan Yang. Lassie: Learning articulated shape from sparse image ensemble via 3d part discovery. In *NeurIPS*, 2022. 1

[1] [4]: `https://github.com/NVlabs/UMR` , [3]: `https://github.com/nileshkulkarni/acsm/`

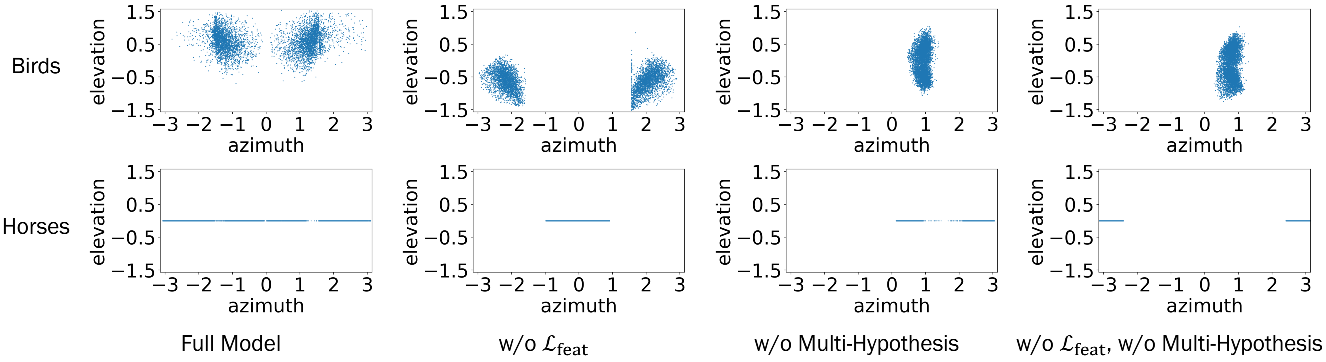

Figure 3. **Visualisations of the Viewpoint Prediction Distributions of the Ablated Models.** We demonstrate that both feature reconstruction loss $\mathcal{L}_{\text{feat}}$ and multi-hypothesis viewpoint prediction are needed to successfully recover a full range of viewpoints. The viewpoint prediction collapses to a limited range as demonstrated by its azimuth without these two components. Note that for Horses, we only predict the azimuth of the viewpoint, as most of the horse images were taken with little elevation.

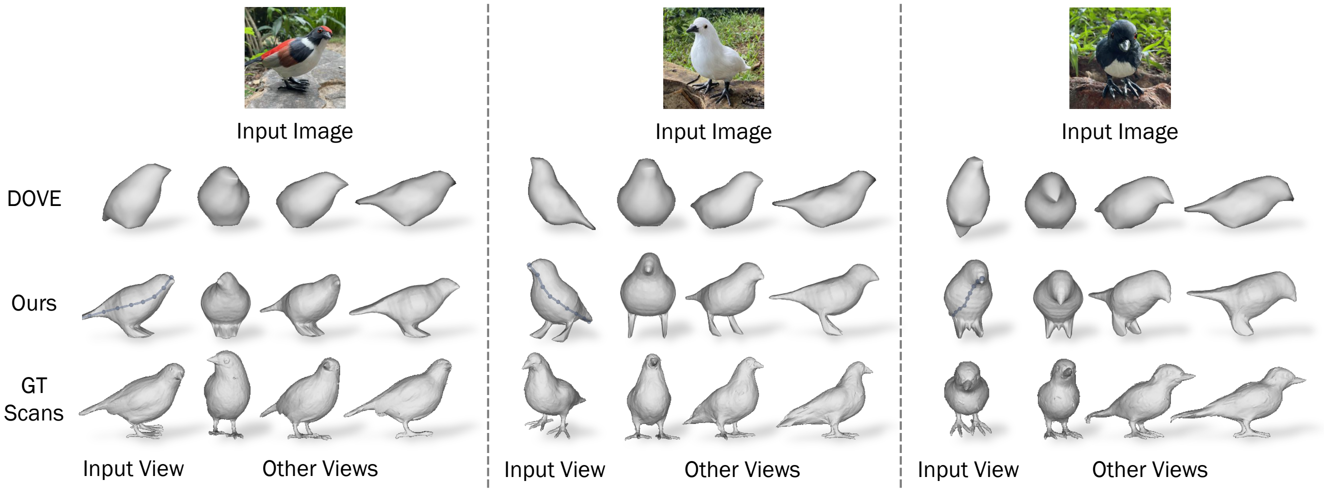

Figure 4. **Visual Comparison on Toy Bird Scans Evaluations.** We compare the reconstructed shapes with scanned ground-truth shapes from Toy Bird Scans dataset. We show the reconstructed mesh from the input view and three additional views. Our model is able to predict finer shape details including the bird's legs as opposed to the prior work of DOVE [6].

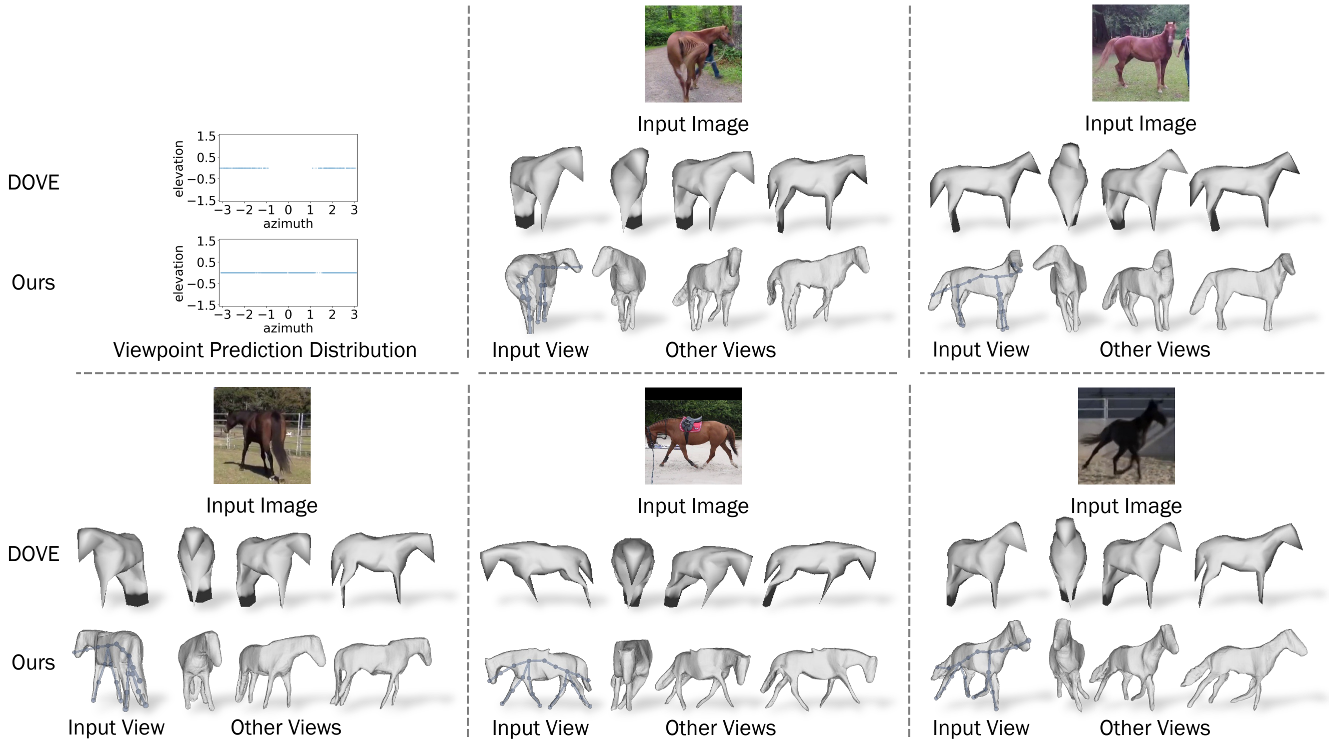

Figure 5. **Comparison with DOVE [6] on Horses.** We visualise the distribution of predicted viewpoints on the test set together with additional qualitative results. Our method is able to recover the full range viewpoint azimuth, while DOVE covers only a portion of possible azimuths. This is further illustrated by the qualitative results, where DOVE often fails to predict the correct viewpoint as opposed to our method. Moreover, our predicted shape is far more detailed. Note that for horses, we only predict azimuth rotations, as most of the horse images were taken with little elevation.

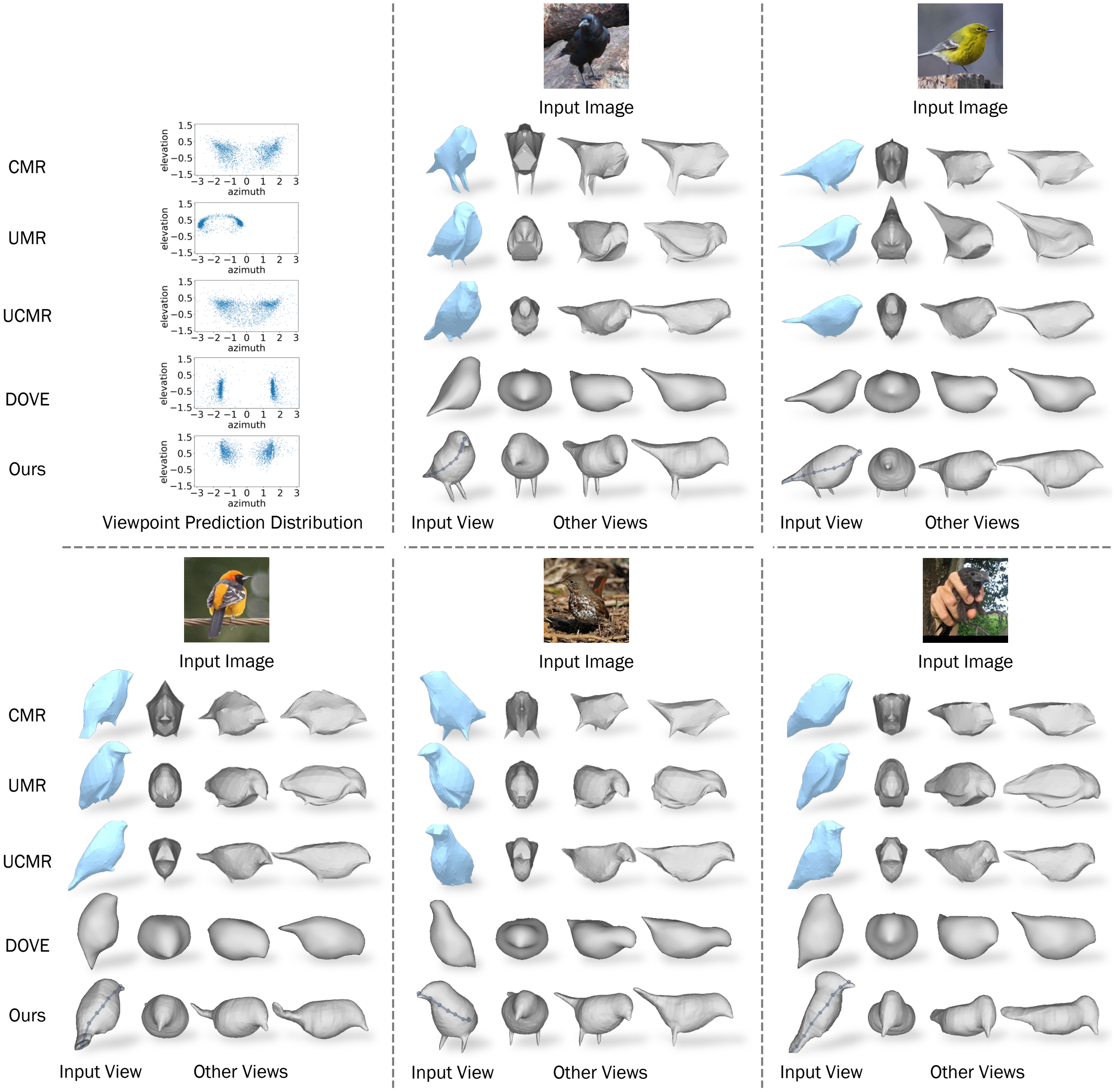

Figure 6. **Comparison with Previous Methods on Horses.** As in Fig. 5 we visualise the distribution of predicted viewpoints on the test set together with additional qualitative results. The plot of viewpoint prediction distribution on CUB test set shows that our method is able to recover a wide range of viewpoints while UMR, which uses a similar level of supervision, is able to predict only frontal poses. We also present additional qualitative results on CUB test set demonstrating that our method recovers shapes with greater details than previous works while using significantly less supervision.

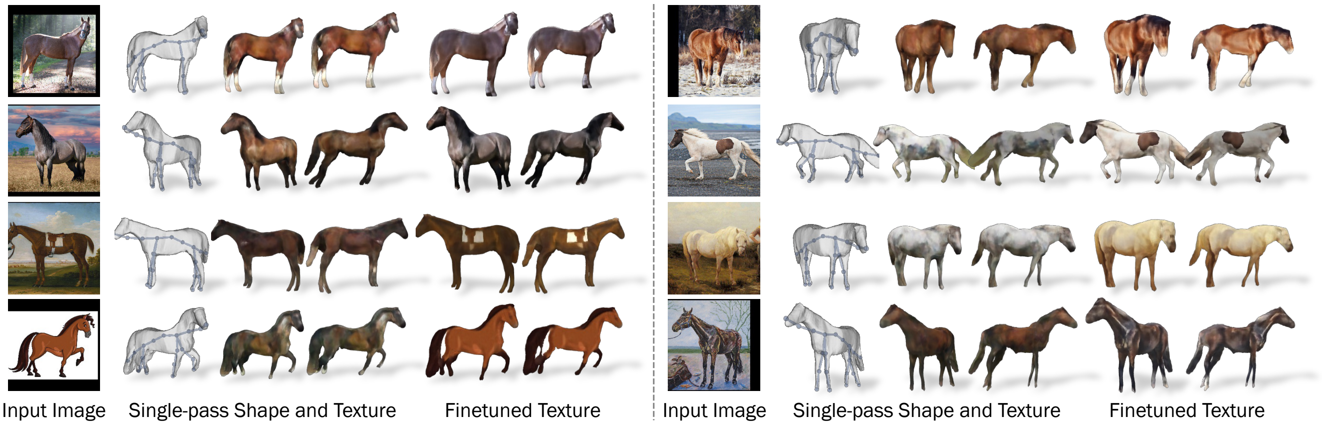

Figure 7. **Texture Finetuning at Test Time.** We show a shape and texture prediction from an input view and one additional view together with a finetuned version of the texture. We demonstrate that a simple finetuning of the texture on the input image can produce high-quality textures for images that are too far from the training set distribution.

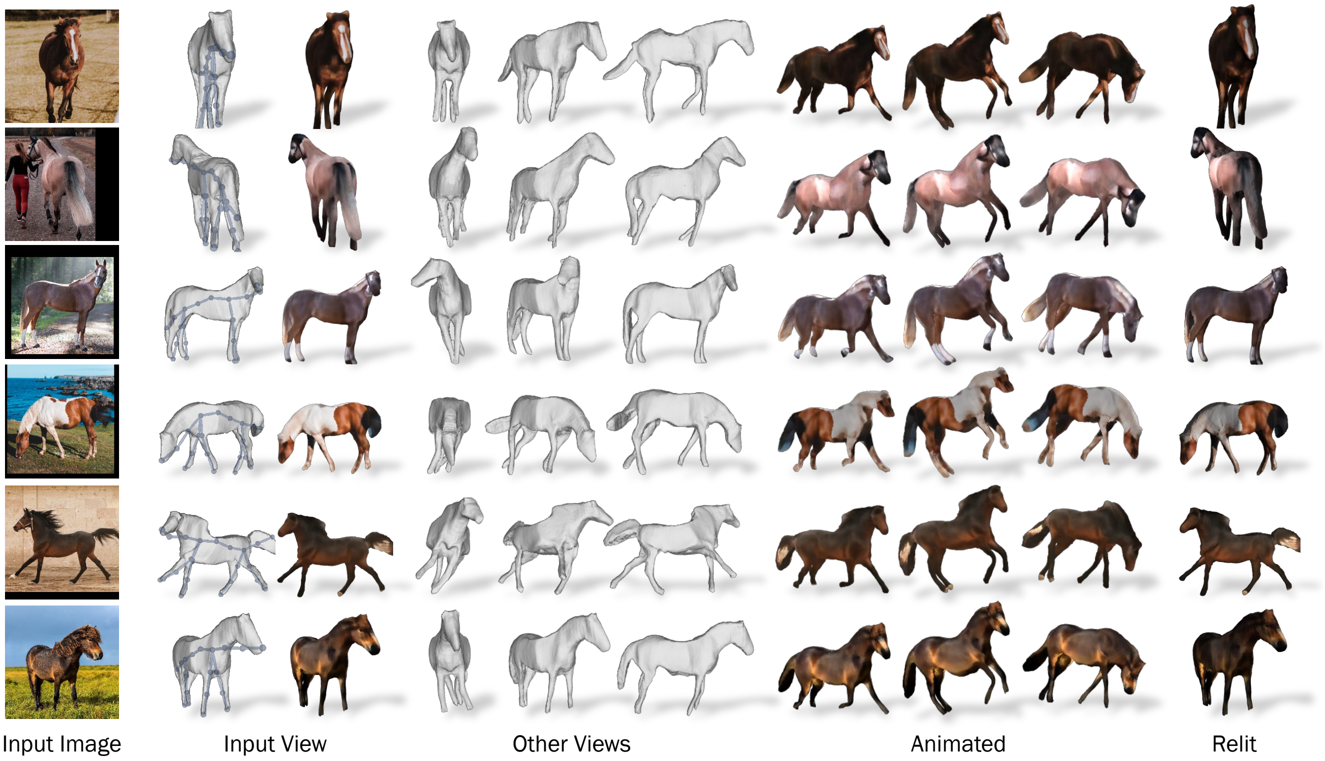

Figure 8. **Reconstruction of Real Horse Images.** We show the predicted mesh from the input view and three additional views. We also demonstrate that our shape can be animated by articulating the estimated skeleton. Finally, as our method decomposes albedo and lightning, our predictions can be easily relit.

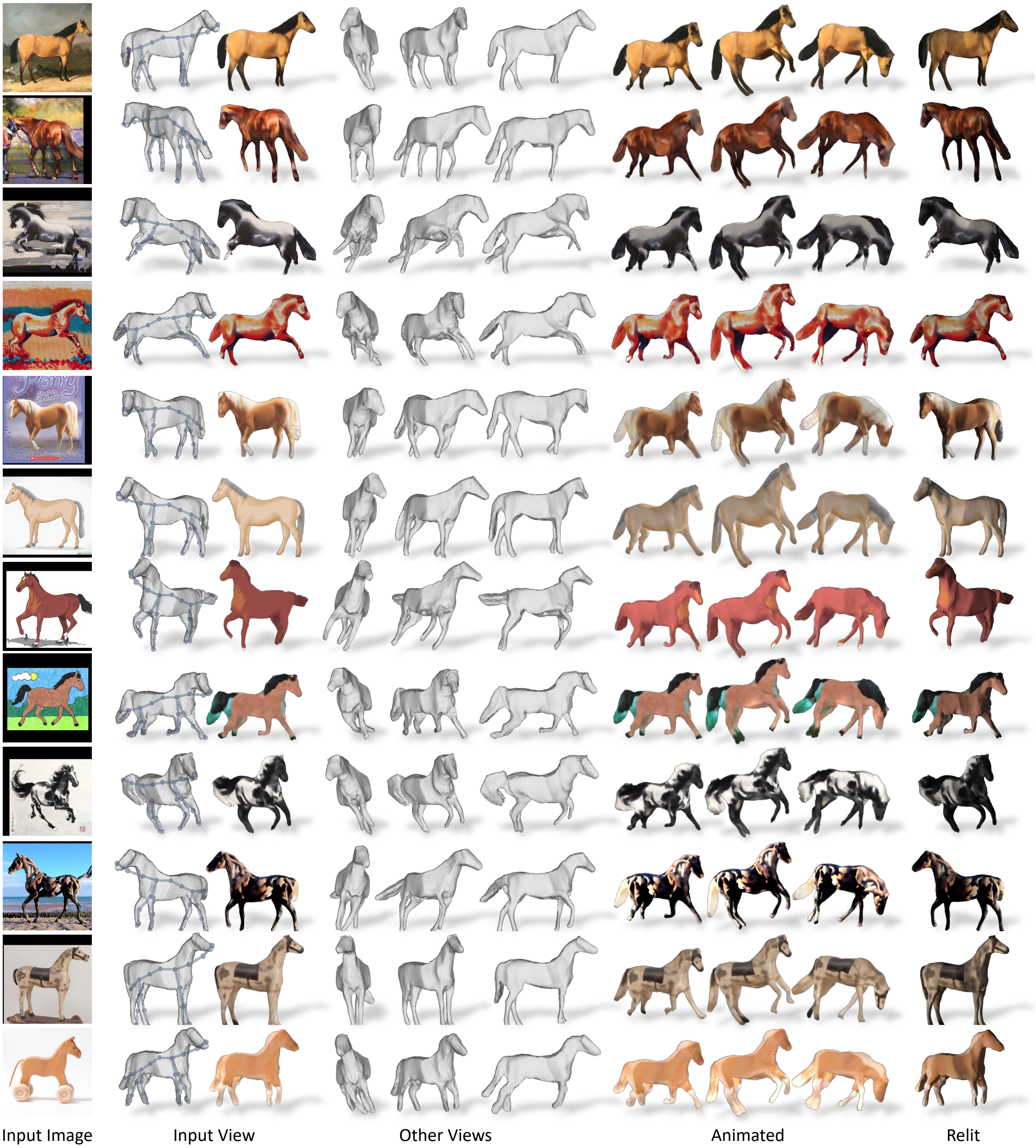

Figure 9. **Reconstruction of Abstract Horse Drawings and Artefacts.** As in Fig. 8, here we show the predicted meshes from the input view and three additional views together with the animated and relit versions. The results demonstrate excellent generalisation of our method on images far from the distribution of the training set which consists only of real horse images.

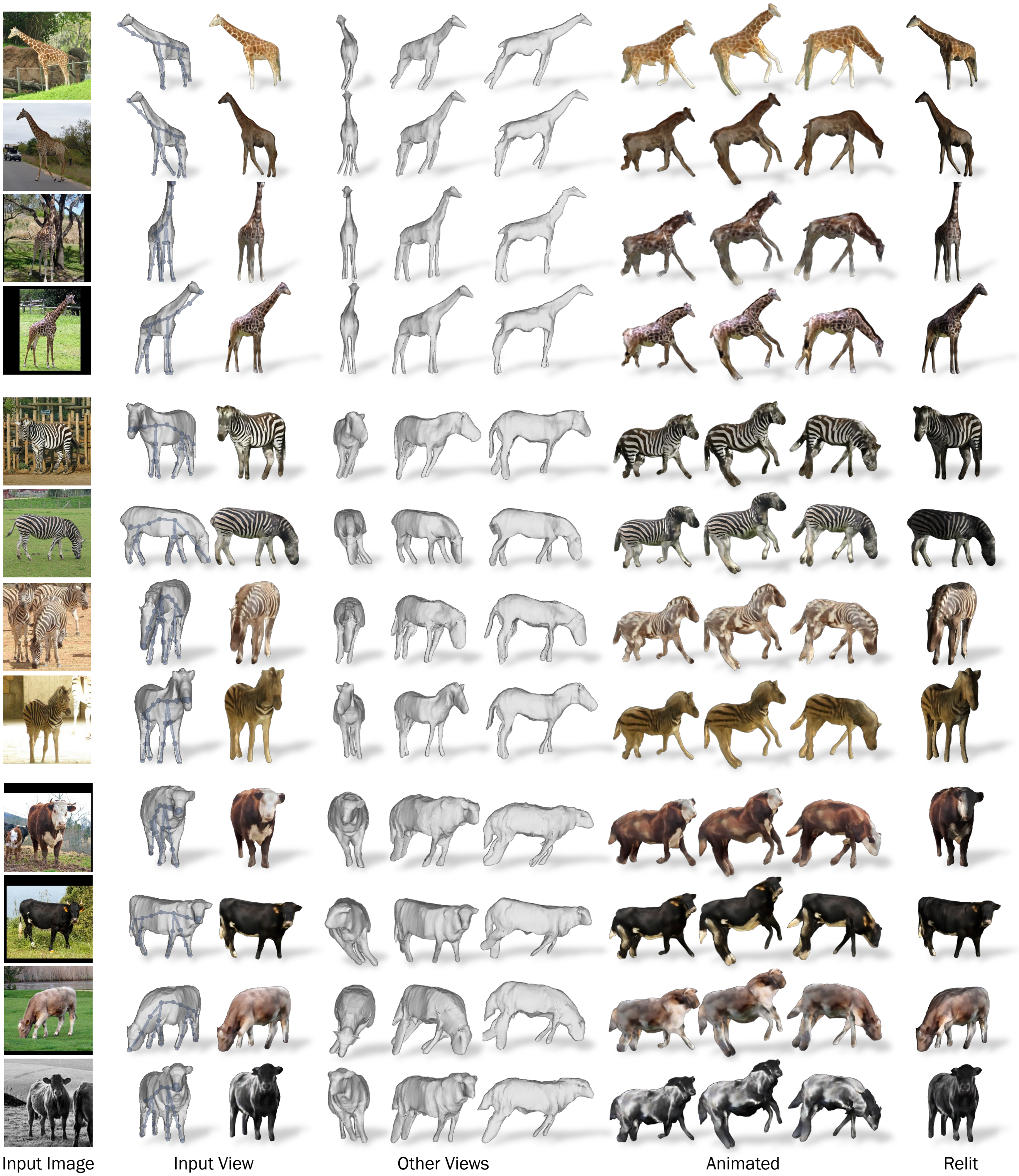

Figure 10. **Reconstruction of Giraffes, Zebras and Cows.** After finetuning on new categories, our method generalises to various animal classes with highly different underlying shapes. We show the predicted mesh from the input view and three additional views together with animated versions of the shape obtained by articulating the estimated skeleton. Finally, we show a relit version.