arXiv:2211.12497v1 [cs.CV] 22 Nov 2022

MagicPony: Learning Articulated 3D Animals in the Wild

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

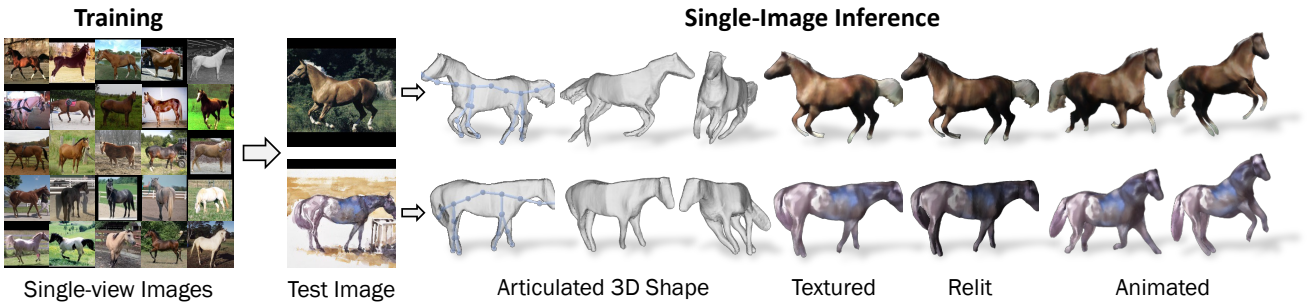

Figure 1. **Learning Articulated 3D Animals in the Wild.** Our method trains on a collection of single-view images of an animal category and produces a model that can predict an articulated 3D shape of the instance from a single input image, which can be animated and relit.

Abstract

We consider the problem of learning a function that can estimate the 3D shape, articulation, viewpoint, texture, and lighting of an articulated animal like a horse, given a single test image. We present a new method, dubbed MagicPony, that learns this function purely from in-the-wild single-view images of the object category, with minimal assumptions about the topology of deformation. At its core is an implicit-explicit representation of articulated shape and appearance, combining the strengths of neural fields and meshes. In order to help the model understand an object's shape and pose, we distil the knowledge captured by an off-the-shelf self-supervised vision transformer and fuse it into the 3D model. To overcome common local optima in viewpoint estimation, we further introduce a new viewpoint sampling scheme that comes at no added training cost. Compared to prior works, we show significant quantitative and qualitative improvements on this challenging task. The model also demonstrates excellent generalisation in reconstructing abstract drawings and artefacts, despite the fact that it is only trained on real images.

1. Introduction

A model that can reconstruct the 3D shape of a deformable object from a single image must know *a priori* what the possible shapes and appearances of the object are. Learning such a prior usually requires ad-hoc data acquisition setups [2, 19, 35, 36], involving at least multiple cameras, and often laser scanners, domes and other hardware, not to mention manual labour. This is viable for specific objects such as humans that are of particular interest in applications, but it is unlikely to scale to the long tail of objects that can appear in natural images.

The alternative is to learn the necessary 3D priors from what is available in abundance, namely 2D images. Unfortunately, for any given type of object, the space of possible 3D reconstructions is characterised by rich and yet highly-structured shape, pose and appearance variations across instances. This prior must then be *learned while using it to reconstruct* the available 2D data *in 3D*, which is a major challenge. This is especially true when no other mode of manual or automated supervision is given.

In this paper, we propose a novel approach to learning 3D models of articulated object categories such as horses and birds with only *single-view* input images for training, which we dub **MagicPony**. We leverage recent progress in unsupervised representation learning, unsupervised image matching, efficient implicit-explicit shape representations and neural rendering, to devise a new auto-encoder architecture that reconstructs the 3D shape, articulation and texture of each object instance from a single test image. For training, we only require a 2D segmenter for the objects and a description of the topology and symmetry of the 3D skeleton (*e.g.*, the number of legs). We *do not* re-

*Equal contribution.

quire a-priori knowledge of 3D shape models, keypoints, viewpoints or any other 2D or 3D cue often used in prior work [14, 20, 21, 30]. From this, we learn a function that, at test time, can estimate the shape and texture of a new object from a single image, in a feed-forward manner. The function exhibits remarkable generalisation properties, capable of reconstructing objects in *abstract drawings*, despite being trained on real images only.

We identify a small number of key challenges that must be addressed to solve this problem, and propose new effective solutions to each of them.

The first problem is viewpoint estimation. Given only raw 2D images of a given object type, until a 3D model is available it is very difficult to reliably assign the images to different aspects or viewpoints of the object. Prior works have shown that this task can be simplified significantly if at least 2D point correspondences between images are available [20, 21, 30]. To avoid requiring keypoint supervision, we derive reasonable but noisy correspondences by *fusing* into the 3D model knowledge distilled from DINO-ViT [5], a self-supervised visual transformer network (ViT) [27]. We add to that a new efficient disambiguation scheme that explores multiple viewpoint assignment hypotheses at essentially no cost, avoiding local optima that are caused by greedily matching the noisy 2D correspondences.

The second problem is how to represent the 3D shape, appearance and deformations of the object. Most prior works have used textured meshes [13, 14, 20, 30, 31, 68, 72], but these are difficult to optimise from scratch, leading to problems that often require ad-hoc heuristics such as re-meshing [13, 72]. The other and increasingly popular approach is to use a volumetric representation such as a neural radiance field [3, 41, 53, 64]. These representations can model complex shapes that evolve in a seamless manner during training, including changing their topology. However, this modelling freedom comes at the cost of over-parameterisation, which is particularly problematic in monocular reconstruction and often leads to meaningless short-cut solutions [64]. Furthermore, modelling articulation with a volumetric representation is difficult. Deformations are naturally defined only for the object's surface and its interior and from the canonical/pose-free space to the posed space. However, ray marching requires extending these transformations around the object (which is not 'physical') and often needs to invert them (which is hard [8]).

We address these issues by using a hybrid volumetric-mesh representation, based on DMTet [57]. Shape and appearance are defined volumetrically in canonical space, but a mesh is extracted on the fly for posing and rendering. This sidesteps the challenges of using neural rendering directly, while retaining most of its advantages and enabling the use of powerful shape regularisers. We compare our method to prior work on two challenging articulated object categories: horses and birds, and show that our method obtains *significantly better* quantitative and qualitative results while using *significantly less* supervision.

Table 1. **Related Work Overview on Weakly-supervised Learning of 3D Objects.** Annotations: template shape, viewpoint, 2D keypoint, object mask, optical flow, video, [1]coarse template shape from keypoints, [2]camera estimated from keypoints using SfM, [3]outputs texture flow, [4]shape bases initialised from CMR. [†]UMR relies on part segmentations from SCOPS [18].

	Supervision						Output				
Method							3D	2.5D	Motion	View	Texture
Unsup3D [70]								✓		✓	✓
CSM [29]	✓			✓						✓	
A-CSM [28]	✓			✓					✓	✓	
CMR [20]	(✓)[1]	(✓)[2]	✓	✓			✓			✓	(✓)[3]
U-CMR [14]	✓			✓			✓			✓	✓
UMR[†] [31]				✓			✓			✓	(✓)[3]
ACMR [30]	(✓)[4]	(✓)[2]	✓	✓			✓			✓	(✓)[3]
DOVE [68]				✓	✓	✓	✓		✓	✓	✓
Ours				✓			✓		✓	✓	✓

To summarise, we make the following **contributions**: (1) A new 3D object learning framework that combines recent advances in unsupervised learning, 3D representations and neural rendering, achieving better reconstruction results with less supervision; (2) An effective mechanism for fusing self-supervised features from DINO-ViT into the 3D model as a form of self-supervision; (3) An efficient multi-hypothesis viewpoint prediction scheme that avoids local optima in reconstruction with no additional cost; (4) Extensive quantitative and qualitative evaluations, including demonstrating generalisation to abstract drawings.

2. Related Work

Weakly-supervised Learning of 3D Objects. Most related to our work are weakly-supervised methods for single-image 3D object reconstruction that learn from a collection of images or videos of an object category [14, 17, 20, 22, 23, 25, 26, 30, 31, 34, 66, 68, 69]. Due to the inherent ambiguity of the problem, these methods usually rely on heavy geometric supervision (see Tab. 1) in addtion to object masks, such as 2D keypoints [20, 30], template shapes [14, 28, 29]. UMR [31] forgoes the requirement of keypoints and template shapes by leveraging weakly-supervised object part segmentations (SCOPS [18]), but produces coarse symmetric shapes similar to CMR [20]. DOVE [68] learns coarse articulated objects by exploiting temporal information from videos with optical flow supervision. All these methods require object masks for supervision, except Unsup3D [70] which exploits bilateral symmetry for reconstructing roughly frontal objects, like faces, and UNICORN [42] which uses a progressive conditioning strategy with a heavily constrained bottleneck, leading

to coarse reconstructions. Another emerging paradigm is to leverage generative models [7, 46, 50, 56] that encourages images rendered from viewpoints sampled from a prior distribution to be realistic. These models, however, often rely heavily on an accurate estimation of viewpoint distribution and/or a powerful 2D image generative model, both of which are difficult to obtain in small-data scenarios.

Optimization from Multi-views and Videos. 3D reconstruction traditionally relies on epipolar geometry of multi-view images of static scenes [11, 16]. Neural Radiance Fields (NeRF) [3, 41, 53, 64] have recently emerged as a powerful volumetric representation for multi-view reconstruction given accurate cameras. A recent line of work, LASR [72], ViSER [73] and BANMo [74], optimises 3D shapes of deformable objects from a small set of monocular videos, with heavily engineered optimisation strategies using optical flow and mask supervision, and additionally DensePose [45] annotations for highly deformable animals. Concurrent work of LASSIE [75] also leverages DINO-ViT [5] image features for supervision, but optimises a part-based model on a small collection of images ($\sim$30) of an object category. It uses a shared shape model for all instances with only per-instance articulation.

Learnable Deformable 3D Representations. A common 3D representation for weakly-supervised shape learning is triangular meshes. Deformation on meshes can be modelled by estimating offsets of individual vertices [20, 30, 68], which typically requires heavy regularisation, such as As-Rigid-As-Possible (ARAP) [58]. To constrain the space of deformations, many works often utilise lower-dimensional models, such as cages [76] or linear blend skinning controlled with skeletal bones [36, 68] or Gaussian control points [72, 73]. Parametric models, like SMPL [36] and SMAL [78], allow for realistic control of the shape through learned parameters, but often require an extensive collection of 3D scans to learn from.

Dynamic neural fields [12, 47, 49, 51, 52, 59, 60, 74], such as D-NeRF [51], extend NeRF [41] with time-varying components. A-NeRF [59] proposes a radiance field conditioned on an articulated skeleton, while BANMo [74] uses learned Gaussian control points similarly to [72, 73]. One key limitation of these implicit representations is the requirement of an inverse transformation from 3D world coordinates back to the canonical space, which is often harder to learn compared to forward deformation [8]. We propose a hybrid SDF-mesh representation, extending DMTet [57] with an articulation mechanism, which combines the expressiveness of implicit models with the simplicity of mesh deformation and articulation.

Viewpoint Prediction in the Wild. Viewpoint prediction in weakly-supervised settings is challenging, as it is prone to local optima induced by projection and common object symmetries. Existing solutions are based around learning a distribution over multiple possible viewpoints. U-CMR [14] proposes a camera-multiplex that optimises over 40 cameras for each training sample using a given template shape, which is expensive as evaluating each hypothesis involves a rendering step. We propose a new viewpoint exploration scheme that requires evaluating *only* one hypothesis in each step with essentially no added training cost, and does not rely on a given template shape but a jointly learned model. Recent work of Implicit-PDF [44] proposes to learn viewpoint distributions implicitly by estimating a probability for each image-pose pair of a fixed object instance. RelPose [77] extends this to learning relative poses of image pairs. Our model also predicts a probability associated with each viewpoint hypothesis, but instead relates an instance to a learned category-level prior model.

3. Method

Given only a collection of single-view images of a deformable object category collected in the wild, our goal is to train a monocular reconstruction network that, at test time, predicts the 3D shape, articulated pose, albedo and lighting from a single image of the object.

During training, our method requires no geometric supervision other than the images, 2D foreground segmentations of the object obtained automatically from an off-the-shelf method like [24], and a description of the topology of the object's skeleton (*e.g.*, the number of legs). It is based on several key ideas: an implicit-explicit shape representation (Sec. 3.1), a carefully designed hierarchy of shapes from generic category-specific prior to articulated instances (Sec. 3.2), and an effective viewpoint prediction scheme that leverages self-supervised correspondences and an efficient multi-hypothesis exploration scheme to avoid local minima (Sec. 3.3), as illustrated in Fig. 2.

3.1. Implicit-Explicit 3D Shape Representation

The choice of representation for the 3D shape of the object is critical. It must be (1) sufficiently *expressive* for modelling fine-grained shape details and deformations and (2) *effective* for learning with only weak supervision through image reconstruction.

A common choice is to adopt an explicit representation like a triangular mesh [20, 68, 72]. Meshes are conceptually simple and more compact than volumetric representations; they provide access to surface properties such as smoothness, which can be used to regularise the shape; they can be easily deformed by moving their vertices, which simplifies constructing the shape space, *e.g.* by using blend skinning. While meshes have been found to be an effective tool to learn shapes and deformations in weakly-supervised conditions, they are limited by their inherent difficulty in modelling topology changes, which may be necessary as the

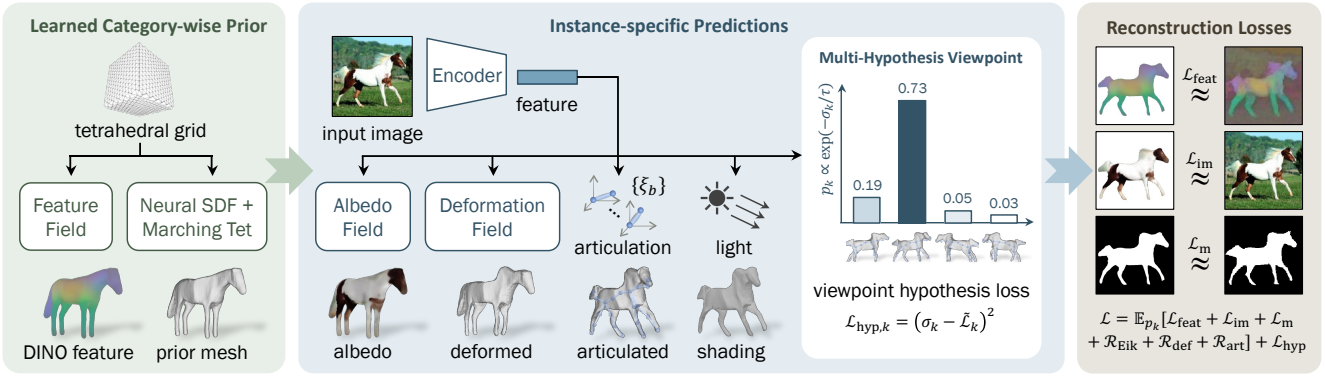

Figure 2. **Training Pipeline.** Given a collection of single-view images of an animal category, our model learns a category-specific prior shape using an implicit-explicit representation, together with a feature field that allows us to fuse self-supervised correspondences through a feature rendering loss. This prior shape is deformed, articulated and shaded based on features extracted from a training image. To combat local minima in viewpoint prediction, we introduce an efficient scheme that explores multiple hypotheses at essentially no additional cost. The entire pipeline is trained with reconstruction losses end-to-end, except for the frozen image encoder pre-trained via self-supervision.

model evolves during training. They can also develop defects such as folds, requiring occasional remeshing [13, 72].

The alternative to meshes are implicit representations like SDFs and radiance fields [40, 41, 48, 67], which naturally allow for topological changes during training. However, due to their high capacity, these representations require substantial supervision, usually in the form of a dense coverage of the different object viewpoints [41, 67] or 3D ground-truth [40, 48]. Furthermore, modelling deformations with implicit representations is harder than for explicit meshes, so they have found more success in modelling single static scenes rather than a family of different objects.

Here we adopt a representation that combines the advantages of both: the shape of the object is represented implicitly by a neural field but converted on-the-fly into an explicit mesh via a fast marching tetrahedra method of [43, 57], a relative of cubes-like [32]. The object's shape $S = \{\mathbf{x} \in \mathbb{R}^3 | s(\mathbf{x}) = 0\}$ is defined as the isosurface of the signed distance function (SDF) $s : \mathbb{R}^3 \mapsto \mathbb{R}$. We implement the SDF s using a Multi-Layer Perceptron (MLP) taking 3D coordinates as input. To encourage s to be a signed distance function, we minimise the Eikonal regulariser [15]:

$$\mathcal{R}_{\text{Eik}} = \sum_{x \in \mathbb{R}^3} (\|\nabla s(\mathbf{x})\|_2 - 1)^2. \tag{1}$$

We use DMTet's implementation [57] of Differentiable Marching Tetrahedra [9], which extracts a watertight mesh M from s on the fly.

3.2. 3D Shape Modelling and Prediction

In order to represent the range of possible shapes of an object category, we use a *hierarchical* model, starting from a category-wise template shape, deforming the latter to model the shape of a specific object instance, and finally matching the pose of the object to a specific image using blend skinning (Fig. 2). These deformations are predicted from a single image by learning networks that share a common image encoder. We discuss these components below.

Image Encoder. The networks that predict the object shape and pose from a single image solve a difficult task and their design is critical. To facilitate their job, we encode the image by means of a self-supervised Vision Transformer (ViT [27]) like DINO-ViT [5]. The ViT extracts keys and output tokens from the image, which were respectively found to code for semantic correspondences [1] and image appearance [62]. Hence, we use keys to predict the object shape and pose and output tokens to infer its appearance.

Formally, the ViT Φ divides the input image $I \in \mathbb{R}^{3\times H\times W}$ into a grid of patches and uses them as tokens in a sequence of attention layers. We take the *keys* $\Phi_{\text{k}} \in \mathbb{R}^{D\times H_p\times W_p}$ and the *output tokens* $\Phi_{\text{o}} \in \mathbb{R}^{D\times H_p\times W_p}$ from the last attention layer and feed them into two separate (convolutional) networks to obtain a single global key feature $\phi_{\text{k}} = f_{\text{k}}(\Phi_{\text{k}}) \in \mathbb{R}^D$ and a single global output feature $\phi_{\text{o}} = f_{\text{o}}(\Phi_{\text{o}}) \in \mathbb{R}^D$ for the whole image respectively.

Category-wise Template Shape. We capture the similarities across all instances with a category-wise template, factoring out instance-specific variations. The template shape is represented by the SDF s, initialised with a generic ellipsoid, elongated along the z-axis. The technique of Sec. 3.1 is used to extract a watertight mesh from it, the vertices of which we denote by the symbol V_{pr}.

Instance-specific Shape. The shape of each object instance V_{ins} differs slightly from the template shape V_{pr}. This is accounted for by the deformation: $V_{\text{ins},i} = V_{\text{pr},i} + \Delta V_i$, where the displacement $\Delta V_i = f_{\Delta V}(V_{\text{pr},i}, \phi_{\text{k}})$ is predicted by an MLP $f_{\Delta V}$ from the image features ϕ_k for each vertex $V_{\text{pr},i}$. We exploit the fact that many objects have a bilaterally symmetric structure, and enforce both the prior shape

and instance deformation to be symmetric by mirroring the query locations for the MLPs about the yz-plane. We further restrict the deformation to be modest in magnitude.

Posing. In order to match the 3D shape observed in the image, the instance-specific mesh $V_{\text{ins},i}$ is further deformed by the *posing function* $V_i = g(V_{\text{ins},i}, \xi)$, where ξ are the pose parameters. Posing accounts for the large, but very controlled deformations caused by the underlying skeletal structure of the objects. Attempting to model these directly by the less structured deformation in $V_{\text{ins},i}$, while conceptually possible by relaxing the symmetry and magnitude constraints, would lead to severe overfitting.

We model the pose of the object using blend skinning [6, 38], controlled by a set of bone rotations $\xi_b \in SO(3), b = 2, \dots, B$, and a rigid pose (or viewpoint) $\xi_1 \in SE(3)$. Specifically, we initialise a set of rest-pose bone locations $\mathbf{J}_b$ on the instance mesh using simple heuristics.[1] Each bone b (except the root) has exactly one parent $\pi(b)$, forming a tree structure. Each vertex V_i is associated to the bones by the skinning weights $w_{i,b}$, defined based on the relative proximity to each bone (see Sec. 7.1). The vertices are then posed by the linear blend *skinning equation*:

$$V_i(\xi) = \left(\sum_{b=1}^{B} w_{i,b} G_b(\xi) G_b(\xi^*)^{-1} \right) V_{\text{ins},i}, \qquad G_b = G_{\pi(b)} \circ g_b, \;\; g_b(\xi) = \begin{bmatrix} R_{\xi_b} & \mathbf{J}_b \\ 0 & 1 \end{bmatrix}, \tag{2}$$

where ξ^* denotes the bone rotations at rest pose.

Pose Prediction. Similarly to the instance-specific deformation ΔV_i, the bone rotations $\xi_{2:B}$ are predicted by a network f_{b} from the image features ϕ_{k}. The viewpoint $\xi_1 \in SE(3)$ is predicted separately, given in Sec. 3.3.

Pose articulation depends on estimating local body joints and benefits from using local features. To extract them, we first project the centroid of each bone based on the rest pose and the predicted viewpoint ξ_1:

$$\mathbf{u}_b = \Pi_{\xi_1} \left(\frac{\mathbf{J}_b + \mathbf{J}_{\pi(b)}}{2} \right),$$

and sample the local feature from the patch key feature map $\Phi_{\text{k}}(\mathbf{u}_b)$ at the projected pixel location $\mathbf{u}_b$.

In order to estimate bone rotations given these features, since these parameters are inter-dependent, we use a transformer architecture [63] for f_{b} that operates on the bone descriptors as input tokens $\nu_b = (\phi_{\text{k}}, \Phi_{\text{k}}(\mathbf{u}_b), b, \mathbf{J}_b, \mathbf{u}_b)$ and predicts the bone rotations, parametrized by Euler angles, as output tokens: $(\xi_2, \dots, \xi_B) = f_{\text{b}}(\nu_2, \dots, \nu_B)$.

[1] A chain of bones going through the two farthest end points along z-axis for birds, and for quadrupeds, additionally four legs branching out from the body bone to the lowest point in each quadrant. See Sec. 7.1 .

3.3. Viewpoint Prediction

Estimating the viewpoint is essential in order to learn the object's shape and articulation. Here we leverage the self-supervised DINO-ViT features, which capture noisy semantic correspondences, and a new multi-hypothesis viewpoint prediction scheme that addresses the remaining ambiguity. These steps are explained next.

Fusing Self-Supervised Correspondences. Prior works have often established noisy but explicit model-image correspondences in order to estimate viewpoint. Here, we take a different approach which is simple, efficient and requires a minimal extension to the model. Like N3FF [61], we *fuse* information from the self-supervised ViT features into the 3D model by learning to render them. This is also similar to BANMo [74], but we use self-supervised features instead of supervised DensePose [45] embeddings.

This is done by adding a coordinate network $\psi : \mathbb{R}^3 \mapsto \mathbb{R}^{D'}$ that predicts a field of D' dimensional features in the canonical space, similar to the analogous network for the RGB colours. These canonical features are then *rendered* as an image $\hat{\Phi}_{\text{k}}$ by using the same process used to render RGB colours. They are trained to minimise a corresponding rendering loss: $\mathcal{L}_{\text{feat}} = \|\tilde{M} \odot (\hat{\Phi}_{\text{k}} - \Phi'_{\text{k}})\|_2^2$, where $\tilde{M} = \hat{M} \odot M$ is the intersection of the rendered mask and the ground-truth mask and $\Phi'_{\text{k}} = \texttt{PCA}(\Phi_{\text{k}})$ is a PCA-reduced (to $D' = 16$ principal components) version of the original DINO-ViT features Φ_k for memory efficiency.

Multi-Hypothesis Viewpoint Prediction. Prior works have found that a major challenge in learning the object viewpoint is the existence of multiple local optima in the reconstruction objective. Some have addressed this issue by sampling a large number of hypotheses for the viewpoint [14], but this is somewhat cumbersome and slow. Instead, we propose a scheme that explores multiple viewpoints *statistically*, but at each iteration samples a single one, and thus comes "for free".

We hence task the model to predict four viewpoint rotation hypotheses $R_k, k \in \{1, 2, 3, 4\}$, each in one of the four quadrants around the object.[2] The model also predicts a score σ_k for each hypothesis, used to evaluate the probability p_k that hypothesis k is the best of the four options as: $p_k = \frac{\exp(-\sigma'_k)}{\sum_j \exp(-\sigma'_j)}$, where $\sigma'_k = \frac{\sigma_k}{\tau}$ sharpens the distribution using a temperature parameter τ that is gradually decreased during training.

The normal approach for learning σ_k is to sample multiple hypotheses and compare their reconstruction loss to determine which one is better. However, computing the loss is expensive as it requires rendering the model. Instead, we suggest to *sample* a single hypothesis each training it-

[2] For bilaterally symmetric animals, there are often four ambiguous orientations towards each of the four xz-quadrants in the canonical space.

eration[3] and simply learn σ_k to predict the expected reconstruction loss $\tilde{\mathcal{L}}_k$, minimizing the objective:

$$\mathcal{L}_{\text{hyp}} = (\sigma_k - \text{sg}[\tilde{\mathcal{L}}_k])^2, \tag{3}$$

where $\text{sg}[\cdot]$ is the stop-gradient operator.

3.4. Appearance and Shading

The albedo of the object is modelled in canonical space by using a coordinate MLP $a_i = f_{\text{a}}(\mathbf{x}_i, \phi_{\text{o}}) \in \mathbb{R}^3$. A lighting network f_{l} also predicts the dominant light direction $l \in \mathbb{R}^2$ and ambient and diffuse intensities $k_s, k_d \in \mathbb{R}$. We used *differentiable deferred mesh rendering* to render the image. This means that the mesh is projected onto the image by a differentiable renderer, obtaining for each pixel the coordinates of its corresponding 3D point in the canonical space. Then, in a second pass the albedo network is evaluated once per pixel, obtaining the value of the albedo a_i which is shaded to obtain the final pixel colours according to equation $I_i = (k_s + k_d \cdot \max\{0, \langle l, n_i \rangle\}) \cdot a_i$, where n_i is the pixel-wise normal of the *posed* mesh.

3.5. Training Objective

The final training objective is:

$$\begin{aligned} \tilde{\mathcal{L}} &= \lambda_{\text{m}}\mathcal{L}_{\text{m}} + \lambda_{\text{im}}\mathcal{L}_{\text{im}} + \lambda_{\text{f}}\mathcal{L}_{\text{feat}} \\ \mathcal{L} &= \mathbb{E}_{p_k}\left[\tilde{\mathcal{L}} + \lambda_{\text{E}}\mathcal{R}_{\text{Eik}} + \lambda_{\text{d}}\mathcal{R}_{\text{def}} + \lambda_{\text{a}}\mathcal{R}_{\text{art}}\right] + \lambda_{\text{h}}\mathcal{L}_{\text{hyp}}. \end{aligned} \tag{4}$$

This loss includes the main components discussed above plus the following additional objectives, capturing other available cues. Naturally, $\mathcal{L}_{\text{m}} = \|\tilde{M} \odot (\hat{I} - I)\|_1$, where $\tilde{M} = \hat{M} \odot M$ encourages the rendered image to match the input. Here M is the object-mask of the input image and $\hat{M}$ is the (differentiable) mask of the rendering.

The mask loss $\mathcal{L}_{\text{m}} = \|\hat{M} - M\|_2^2 + \lambda_{\text{dt}}(\hat{M} \odot \texttt{dt}(M))$ encourages rendered shape to align with the image mask using a distance transform $\texttt{dt}(\cdot)$ for meaningful gradients.

The feature loss $\mathcal{L}_{\text{feat}} = \|\tilde{M} \odot (\hat{\Phi}'_{\text{k}} - \Phi'_{\text{k}})\|_2^2$, where $\Phi'_{\text{k}} = \texttt{PCA}(\Phi_{\text{k}})$ aligns the mesh rendered with the feature texture to the pre-trained feature encoder.

Finally, we regularise the magnitude of the deformation: $\mathcal{R}_{\text{def}} = \sum_i \|f_{\Delta V}(\mathbf{x}_i)\|_2^2$ and similarly the bone rotations: $\mathcal{R}_{\text{art}} = \sum_b \|\xi_b\|_2^2$.

Due to the supervision of the hypothesis scores with the loss itself, we can avoid evaluating all hypotheses. We sample only one hypothesis k at a time, and evaluate $\mathcal{L} = p_k\mathcal{L}_k$. To improve the sampling efficiency, we sample the viewpoint based on the learned probability distribution p_k with a gradually decreasing temperature τ.

[3] We sample viewpoint $k^* = \text{argmax}_{\text{k}}\, p_k$ 80% of the time, and uniformly at random 20% of the time.

4. Experiments

We conduct extensive experiments on a few animal categories, including horses, giraffes, zebras, cows and birds, and compare against prior work both qualitatively and quantitatively on standard benchmarks. We also show that our model trained on real images generalises to abstract drawings, demonstrating the power of unsupervised learning.

4.1. Data

For horses, we use the horse dataset from DOVE [68] containing 10.1k images extracted from YouTube, and supplement it with 541 additional images from three datasets for diversity: Weizmann Horse Database [4], PASCAL [10] and Horse-10 Dataset [39]. For giraffes, zebras and cows, we source images from Microsoft COCO Dataset [33] and keep the ones with little occlusion. Since these datasets are relatively small, we only use them to finetune the pre-trained horse model. For birds, we combine the DOVE dataset [68] and CUB dataset [65] consisting of 57.9k and 11.7k images respectively. We use an off-the-shelf PointRend [24] detector to obtain segmentation masks, crop around the objects and resize them to 256×256. We follow the original train/test splits from CUB and DOVE, and randomly split the rest, resulting in 11.5k/0.8k horse, 513/57 giraffe, 574/64 zebra, 719/80 cow and 63.9k/10.5k bird images. We additionally collect roughly 100 horse images from the Internet to test generalisation.

4.2. Technical Details

The model is implemented using a total of 8 different neural networks. The feature field ψ, template SDF s, albedo field f_{a} and deformation field $f_{\Delta V}$ are implemented using MLPs, which take in 3D coordinates (concatenated with image features for albedo and deformation networks) as input. The image encoder consists of a ViT-S [27] architecture from DINO [5] and two convolutional encoders f_{k} and f_{o} that fuse the patch features into global feature vectors. We use self-supervised pre-trained DINO-ViT with frozen weights, and only train the convolutional encoders. The lighting network f_{l} is a simple MLP that maps global output feature to a 4-channel vector, and the articulation network f_{b} is a transformer architecture with 4 blocks. We use a separate encoder for the viewpoint network identical to f_{k}, as we find empirically that sharing the encoder tends to make the viewpoint learning unstable. Apart from DINO encoder, all components are trained end-to-end from scratch for 100 and 10 epochs on horses and birds respectively, with the instance deformation and articulation disabled for the first 30 and 2 epochs. The DINO features Φ'_{k} used to compute $\mathcal{L}_{\text{feat}}$ are pre-computed with patch size 8 and stride 4 and masked, following [1], and reduced from a channel size of 384 to 16 using PCA. All technical details and hyperparameters are included in Sec. 7.

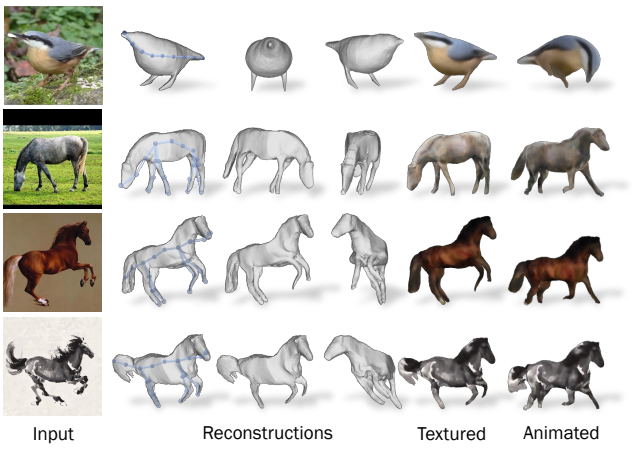

Figure 3. **Single Image Reconstruction.** We show the reconstructed mesh from the input view and two additional views together with predicted texture and animated version of the shape obtained by articulating the estimated skeleton.

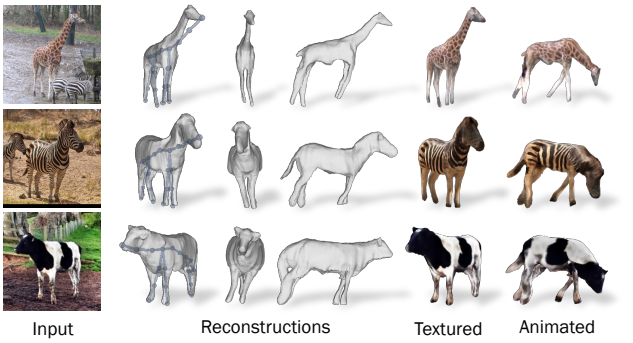

Figure 4. **Reconstructions of Giraffes, Zebras and Cows.** We finetune the horse model on other animal categories, and show that the method generalises well to other animals.

4.3. Qualitative Results

Fig. 3 shows a few reconstructions of horses and birds produced by our model. Given a single 2D image at test time, our model predicts a textured 3D mesh of the object, capturing its fine-grained geometric details, such as the legs and tails of the horses. Our model predicts the articulated pose of the objects, allowing us to easily transfer the pose of one instance from another and animate it in 3D. Although our model is trained on real images only, it demonstrates excellent generalisation to paintings and abstract drawings.

We also show in Fig. 4 that by finetuning the horse model on other animals without any additional modifications (except disabling articulation for the initial 5k iterations), it generalises to various animal categories with highly different shape and appearance, such as giraffes, zebras and cows.

Note that for the examples of abstract horse drawings (last rows in Figs. 1 and 3), since the texture is out of the training distribution, we finetune (only) the albedo network for 100 iterations, which takes less than 5 seconds. This

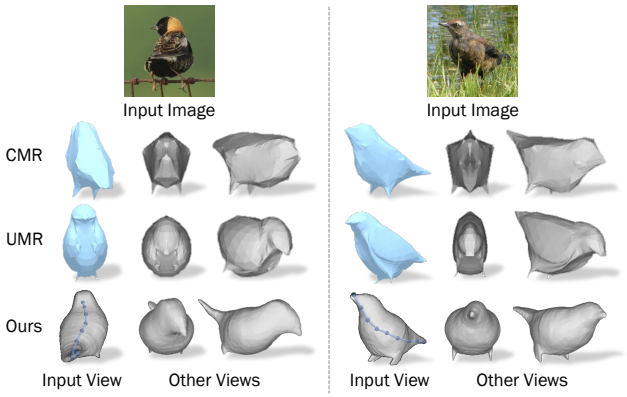

Figure 5. **Comparison on Birds.** Both CMR [20] and UMR [31] often predict inaccurate poses, such as the bird on the left.

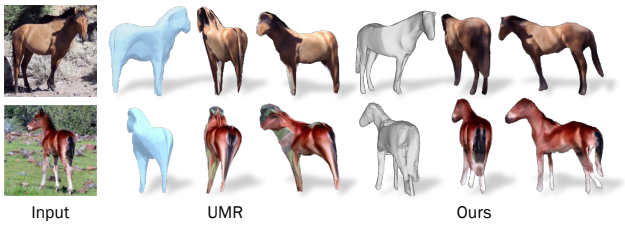

Figure 6. **Comparison with UMR [31] on Horses.** Our model produces much higher quality reconstructions with four legs. UMR is only able to predict symmetrical shapes with two legs.

is also done for giraffes, zebras and cows in Fig. 4, as the training datasets are relatively small for learning complex appearance. Additional qualitative results are presented in Sec. 6 and the *supplementary video*.

4.4. Comparison with Prior Work

We compare with previous weakly-supervised methods on 3D reconstruction of deformable objects. The most relevant prior work is UMR [31] and DOVE [68], as they also only require object masks and weak correspondences from either part segmentations (SCOPS) [18] or video training [68]. Our method leverages *self-supervised* DINO-ViT features, which makes our method the least supervised in this area. We also compare against the well-established baseline of CMR [20], which requires stronger supervision in the form of 2D keypoint annotations and pre-computed cameras, as well as U-CMR [14], which uses a category-specific template shape. To evaluate other methods, we use their code and pre-trained models in public repositories (CMR, UMR, U-CMR) or provided by the authors (DOVE).

Qualitative Comparisons. Figs. 5 and 6 compare our reconstructions against the previous methods on bird and horse images. Since UMR did not release their horse model, we test our model on the examples presented in their paper for side-by-side comparison. Our method predicts more accurate poses and higher quality articulated shapes (as op-

Table 2. **Evaluation on Toy Bird Scans [68].** Baseline results reported by [68]. Annotations: template shape, viewpoint, 2D keypoint, mask, optical flow, video.

	Supervision	Chamfer Distance (cm) ↓
CMR [20]		1.35 ± 0.81
U-CMR [14]		1.82 ± 0.93
UMR [31]	+ SCOPS	1.24 ± 0.75
DOVE [68]		1.51 ± 0.89
Ours	+ DINO	**0.79** ± 0.50

Table 3. **Keypoint Transfer on CUB [65] Dataset.** In addition to the results on the whole CUB test set, we also report results on a subset w/o aquatic birds which none of the methods can sufficiently reconstruct. Our model produces superior results with significantly less supervision.

		PCK@0.1	
Method	Supervision	Entire CUB	w/o aquatic
CMR [20] (from [31])		0.473	-
CMR [20]		0.546	0.591
CMR [20]		0.255	0.277
U-CMR [14]		0.359	0.412
UMR [31]	+ SCOPS	0.512	0.555
DOVE [68]		0.447	0.510
Ours	+ DINO	**0.554**	**0.635**

posed to only symmetric shapes with CMR and UMR). On horses, our method reconstructs all four legs with details while UMR predicts only two. Since UMR copies texture from input image, for fair comparison, we finetune (only) the albedo network for 100 iterations at test time.

Quantitative Comparisons. We evaluate shape reconstruction on the 3D Toy Bird Dataset from DOVE [68], which contains 3D scans of 23 realistic toy bird models paired with 345 photographs of them taken in natural real-world environments. Following [68], we align the predicted mesh with the ground-truth scan using ICP and compute the bi-directional Chamfer Distance between two sets of sample points from the aligned meshes. Tab. 2 summarises the results compared against other methods. Our model produces significantly more accurate reconstructions, resulting in a much lower error (visualisations in Fig. 11).

We also evaluate Keypoint Transfer on the CUB [65] benchmark, a common evaluation metric [20, 31]. We follow the protocol in [31] and sample 10k source and target image pairs. Given a source image, we project all the visible vertices of the predicted mesh onto the image using the predicted viewpoint, and assign each annotated 2D keypoint to its nearest vertex. We then render these vertices using the mesh and viewpoint predicted from a different target image. We measure the error between the projected vertices (corresponding to the transferred keypoints) and the annotated target keypoints using the Percentage of

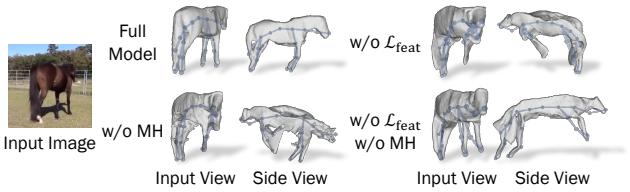

Figure 7. **Ablation Studies.** We train the model without feature loss $\mathcal{L}_{\text{feat}}$, without multi-hypothesis (MH) viewpoint prediction and without both. Without either component, the viewpoint predictions collapse to frontal views, resulting in unnatural shapes.

Correct Keypoints (PCK) metric. As shown in Tab. 3, our model outperforms previous methods, including CMR [20] which requires strong geometric supervision in the form of keypoints. Since CUB dataset contains a wide variety of birds, including aquatic birds like mallards, which are particularly challenging to model due to heavy occlusion and large shape variation, we also report the numbers on a subset of the categories (removing 50 aquatic bird categories), which show even more evident improvements.

4.5. Ablation Studies and Limitations

We present ablation studies on the two key components of the model in Fig. 7, self-supervised feature loss and multi-hypothesis viewpoint prediction, and visualise the learned viewpoint distributions in Fig. 8. Without either component, the learned viewpoints collapse to only frontal views, resulting in unnatural stretched reconstructions.

Although our model reconstructs highly detailed 3D shapes from only a single image and generalises to in-the-wild images, the predicted texture may not preserve sufficient details with a single forward pass and may require additional test-time finetuning. This can potentially be addressed by training on a larger dataset and incorporating recent advances in image generation [54, 55]. Another limitation is the requirement of a pre-defined topology for articulation, which may vary between different animal species. Discovering the articulation structure automatically from raw in-the-wild images will be of great interest for future work. Failure cases are discussed in Sec. 6.5.

5. Conclusions

We have introduced a new model that can learn a 3D model of an articulated object category from single-view images taken in the wild. This model can, at test time, reconstruct the shape, articulation, albedo, and lighting of the object from a single image, and generalises to abstract drawings. Our approach demonstrates the power of combining several recent improvements in self-supervised representation learning together with a new viewpoint sampling scheme. We have shown significantly superior results compared to prior works, even when they use more supervision.

Acknowledgements. We would like to thank Tengda Han, Shu Ishida, Dylan Campbell, Eldar Insafutdinov, Luke Melas-Kyriazi, Ragav Sachdeva and Sagar Vaze for insightful discussions, and Guanqi Zhan and Jaesung Huh for proofreading. Shangzhe Wu is supported by Meta Research. Tomas Jakab is supported by ERC-CoG UNION 101001212. Christian Rupprecht is supported by VisualAI EP/T028572/1 and ERC-CoG UNION 101001212. Andrea Vedaldi is supported by ERC-CoG UNION 101001212.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021. 4, 6

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. *ACM TOG*, 2005. 1

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 2, 3

[4] Eran Borenstein and Shimon Ullman. Class-specific, top-down segmentation. In *ECCV*, 2002. 6

[5] 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. 2, 3, 4, 6

[6] John E Chadwick, David R Haumann, and Richard E Parent. Layered construction for deformable animated characters. *ACM SIGGRAPH Computer Graphics*, 1989. 5

[7] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-GAN: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021. 3

[8] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *ICCV*, 2021. 2, 3

[9] Akio Doi and Akio Koide. An efficient method of triangulating equi-valued surfaces by using tetrahedral cells. *IEICE Transactions on Information and Systems*, 1991. 4

[10] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 6

[11] Olivier Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, 1993. 3

[12] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR*, 2021. 3

[13] Shubham Goel, Georgia Gkioxari, and Jitendra Malik. Differentiable stereopsis: Meshes from multiple views using differentiable rendering. In *CVPR*, 2022. 2, 4

[14] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoints without keypoints. In *ECCV*, 2020. 2, 3, 5, 7, 8

[15] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *ICML*, 2020. 4

[16] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 3

[17] Paul Henderson and Vittorio Ferrari. Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. *IJCV*, 2019. 2

[18] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. SCOPS: Self-supervised co-part segmentation. In *CVPR*, 2019. 2, 7

[19] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 1

[20] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2, 3, 7, 8

[21] Abhishek Kar, Shubham Tulsiani, João Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *CVPR*, 2015. 2

[22] Hiroharu Kato and Tatsuya Harada. Learning view priors for single-view 3d reconstruction. In *CVPR*, 2019. 2

[23] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, 2018. 2

[24] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image segmentation as rendering. In *CVPR*, 2020. 3, 6

[25] Filippos Kokkinos and Iasonas Kokkinos. Learning monocular 3d reconstruction of articulated categories from motion. In *CVPR*, 2021. 2

[26] Filippos Kokkinos and Iasonas Kokkinos. To the point: Correspondence-driven monocular 3d category reconstruction. In *NeurIPS*, 2021. 2

[27] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 4, 6

[28] Nilesh Kulkarni, Abhinav Gupta, David F. Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *CVPR*, 2020. 2

[29] Nilesh Kulkarni, Abhinav Gupta, and Shubham Tulsiani. Canonical surface mapping via geometric cycle consistency. In *ICCV*, 2019. 2

[30] Xueting Li, Sifei Liu, Shalini De Mello, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Online adaptation for consistent mesh reconstruction in the wild. In *NeurIPS*, 2020. 2, 3

[31] 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. 2, 7, 8

[32] Yiyi Liao, Simon Donné, and Andreas Geiger. Deep marching cubes: Learning explicit surface representations. In *CVPR*, 2018. 4

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 6

[34] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, 2019. 2

[35] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *IEEE TPAMI*, 2013. 1

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015. 1, 3

[37] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 12

[38] Nadia Magnenat-Thalmann, E Primeau, and Daniel Thalmann. Abstract muscle action procedures for human face animation. *The Visual Computer*, 1988. 5

[39] Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W. Mathis. Pretraining boosts out-of-domain robustness for pose estimation. In *WACV*, 2021. 6

[40] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 4

[41] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3, 4

[42] Tom Monnier, Matthew Fisher, Alexei A. Efros, and Mathieu Aubry. Share With Thy Neighbors: Single-View Reconstruction by Cross-Instance Consistency. In *ECCV*, 2022. 2

[43] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Mueller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *CVPR*, 2021. 4

[44] Kieran A Murphy, Carlos Esteves, Varun Jampani, Srikumar Ramalingam, and Ameesh Makadia. Implicit-PDF: Non-parametric representation of probability distributions on the rotation manifold. In *ICML*, 2021. 3

[45] Natalia Neverova, David Novotný, and Andrea Vedaldi. Continuous surface embeddings. In *NeurIPS*, 2020. 3, 5

[46] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3d representations from natural images. In *ICCV*, 2019. 3

[47] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, 2021. 3

[48] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, June 2019. 4

[49] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021. 3

[50] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[51] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. In *CVPR*, 2020. 3

[52] Amit Raj, Michael Zollhoefer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. PVA: Pixel-aligned volumetric avatars. In *CVPR*, 2021. 3

[53] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 2, 3

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 8

[55] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 8

[56] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, 2020. 3

[57] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *NeurIPS*, 2021. 2, 3, 4

[58] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, 2007. 3

[59] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-NeRF: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, 2021. 3

[60] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV*, 2021. 3

[61] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural Feature Fusion Fields: 3D distillation of self-supervised 2D image representation. In *3DV*, 2022. 5

[62] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *CVPR*, 2022. 4

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 5

[64] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR*, 2022. 2, 3

[65] 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. 6, 8

[66] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2Mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018. 2

[67] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021. 4

[68] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. DOVE: Learning deformable 3d objects by watching videos. *arXiv preprint arXiv:2107.10844*, 2021. 2, 3, 6, 7, 8, 15

[69] Shangzhe Wu, Ameesh Makadia, Jiajun Wu, Noah Snavely, Richard Tucker, and Angjoo Kanazawa. De-rendering the world's revolutionary artefacts. In *CVPR*, 2021. 2

[70] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3D objects from images in the wild. In *CVPR*, 2020. 2

[71] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 12

[72] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T. Freeman, and Ce Liu. LASR: Learning articulated shape reconstruction from a monocular video. In *CVPR*, 2021. 2, 3, 4

[73] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. ViSER: Video-specific surface embeddings for articulated 3d shape reconstruction. In *NeurIPS*, 2021. 3

[74] Gengshan Yang, Minh Vo, Neverova Natalia, Deva Ramanan, Vedaldi Andrea, and Joo Hanbyul. BANMo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 3, 5

[75] 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. 3

[76] Wang Yifan, Noam Aigerman, Vladimir G Kim, Siddhartha Chaudhuri, and Olga Sorkine-Hornung. Neural cages for detail-preserving 3d deformations. In *CVPR*, 2020. 3

[77] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. RelPose: Predicting probabilistic relative rotation for single objects in the wild. In *ECCV*, 2022. 3

[78] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *CVPR*, 2017. 3

6. Additional Results

6.1. Generalisation to Other Animal Categories

In addition to Fig. 4, we 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.

6.2. Additional Comparisons with Prior Works

Qualitative Comparisons Fig. 12 and Fig. 13 show qualitative comparisons with prior works on horses and birds respectively. Our method is able to predict shapes with finer details and more accurate poses than 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.

Visualisations of the Toy Bird Reconstructions. Supplementary to Tab. 2, we show a qualitative comparison of the predicted shapes and the scanned ground-truth shapes on the Toy Bird Scan dataset in Fig. 11.

6.3. Ablation Studies

In addition to Fig. 7, we examine the effects of both the feature rendering loss $\mathcal{L}_{\text{feat}}$ and the multi-hypothesis viewpoint prediction in Fig. 8, demonstrating that both of the components are essential to prevent the collapse of viewpoint prediction.

6.4. Texture Finetuning

Fig. 14 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 single-pass predictions.

6.5. 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 5 seconds) can remedy this shortcoming. Fig. 14 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. 9. 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. 15. 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.

6.6. Additional Qualitative Results

Additional results of single image reconstruction, animation and relighting can be found in Fig. 15 and the supplementary video. More generalisation results on abstract horse drawings, sculptures and toys are presented in Fig. 16, showing that the model has learned to estimate shape, pose and articulation sufficiently robustly to generalise beyond the training distribution.

7. Additional Technical Details

7.1. 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 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 further add 4 sets of bones for modelling the legs. 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), 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:

$$w_{i,b} = \frac{e^{-d_{i,b}/\tau_{\text{s}}}}{\sum_{k=1}^{B} e^{-d_{i,k}/\tau_{\text{s}}}}, \quad \text{where} \quad d_{i,b} = \min_{r\in[0,1]} \|V_{\text{ins},i} - r\mathbf{J}_b - (1-r)\mathbf{J}_{\pi(b)}\|_2^2 \tag{5}$$

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

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 $(-18°, 18°)$.

7.2. 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. 7.3. 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 [71] with n groups
- LReLU(p): leaky ReLU [37] with a slope p

7.3. Hyper-parameters and Training Details

All hyper-parameters are listed in Tab. 4. 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.

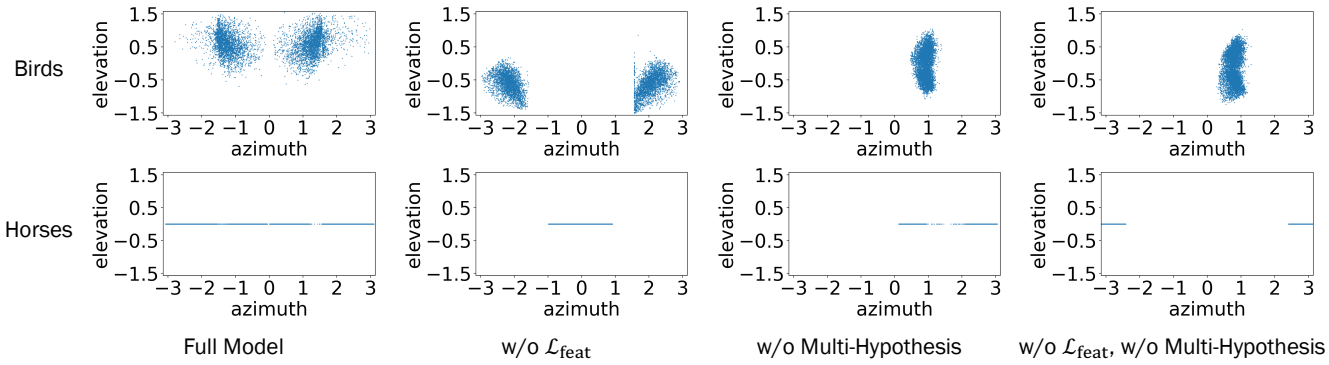

Figure 8. **Visualisations of the Viewpoint Prediction Distributions of the Ablated Models.** We demonstrate that both feature reconstruction loss $\mathcal{L}_{\text{feat}}$ and multi-hypothesis viewpoint prediction are needed to successfully recover a full range of viewpoints. The viewpoint prediction 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.

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	$(0.5, 1.0)$

Table 4. Training details and hyper-parameter settings.

Table 5. Architecture of the encoders f_{k}, f_{o} and the viewpoint network f_{v}.

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) $\rightarrow$ output	1×1

Figure 9. **Incorrect Viewpoint Predictions.** The viewpoint prediction can be less reliable in the case of more extreme input views.

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.

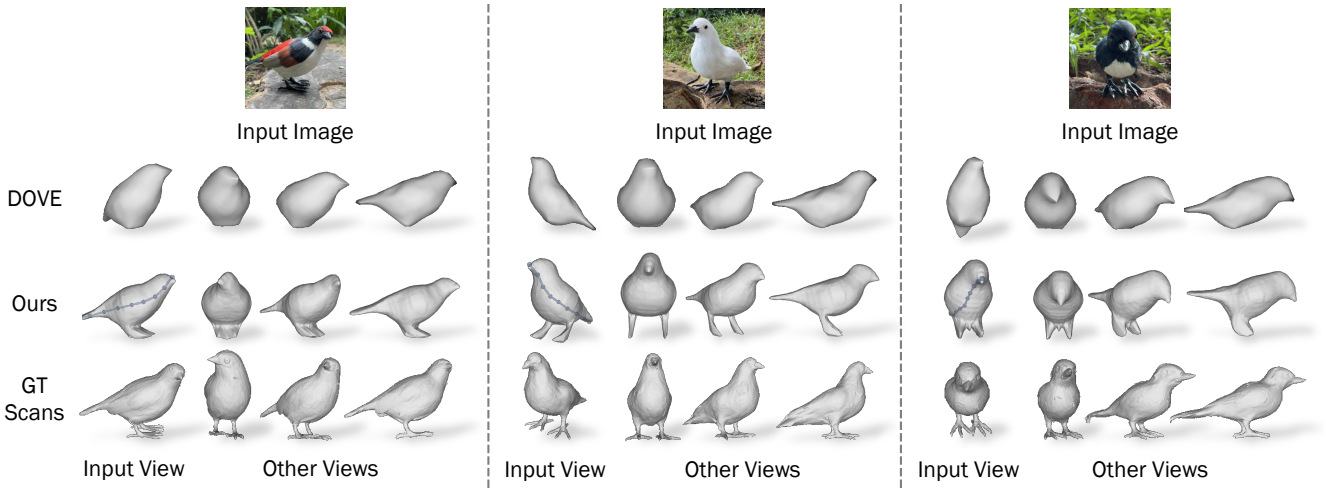

Figure 11. **Visual Comparison on Toy Bird Scans Evaluations.** We compare the reconstructed shapes with scanned ground-truth shapes from Toy Bird Scans dataset. We show the reconstructed mesh from the input view and three additional views. Our model is able to predict finer shape details including the bird's legs as opposed to the prior work of DOVE [68].

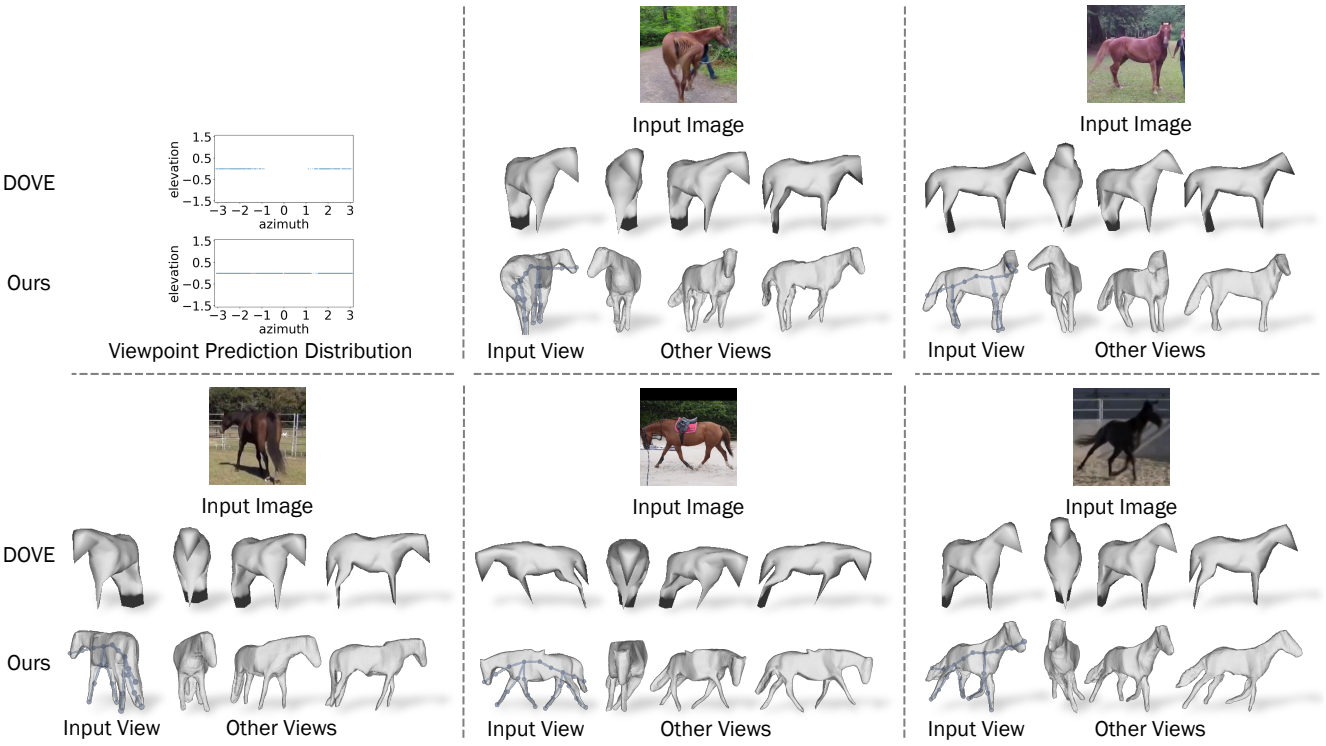

Figure 12. **Comparison with DOVE [68] on Horses.** We visualise the distribution of predicted viewpoints on the test set together with additional qualitative results. Our method is able to recover the full range viewpoint azimuth, while DOVE covers only a portion of possible azimuths. This is further 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.

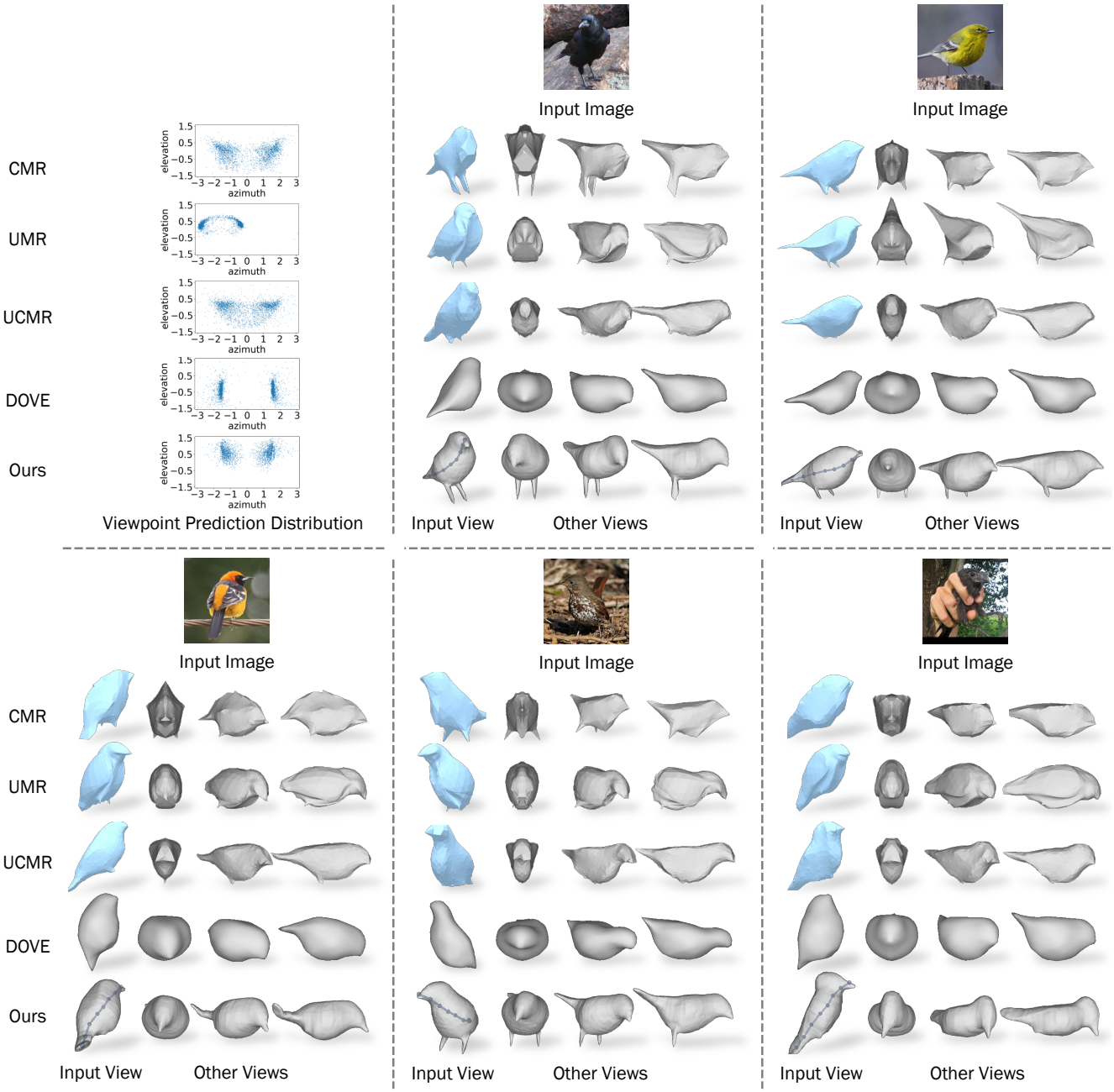

Figure 13. **Comparison with Previous Methods on Horses.** As in Fig. 12 we visualise the distribution of predicted viewpoints on the test set together with additional qualitative results. The plot of viewpoint prediction distribution on CUB test set shows that our method is able to recover a wide range 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.

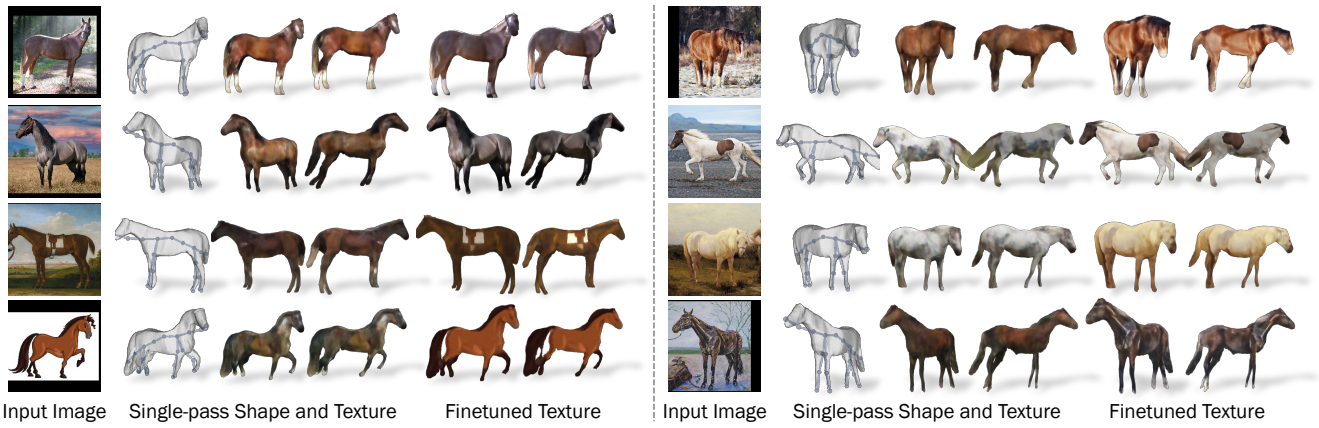

Figure 14. **Texture Finetuning at Test Time.** We show a shape and texture prediction from an input view and one additional view together with a finetuned version of the texture. We demonstrate that a simple finetuning of the texture on the input image can produce high-quality textures for images that are too far from the training set distribution.

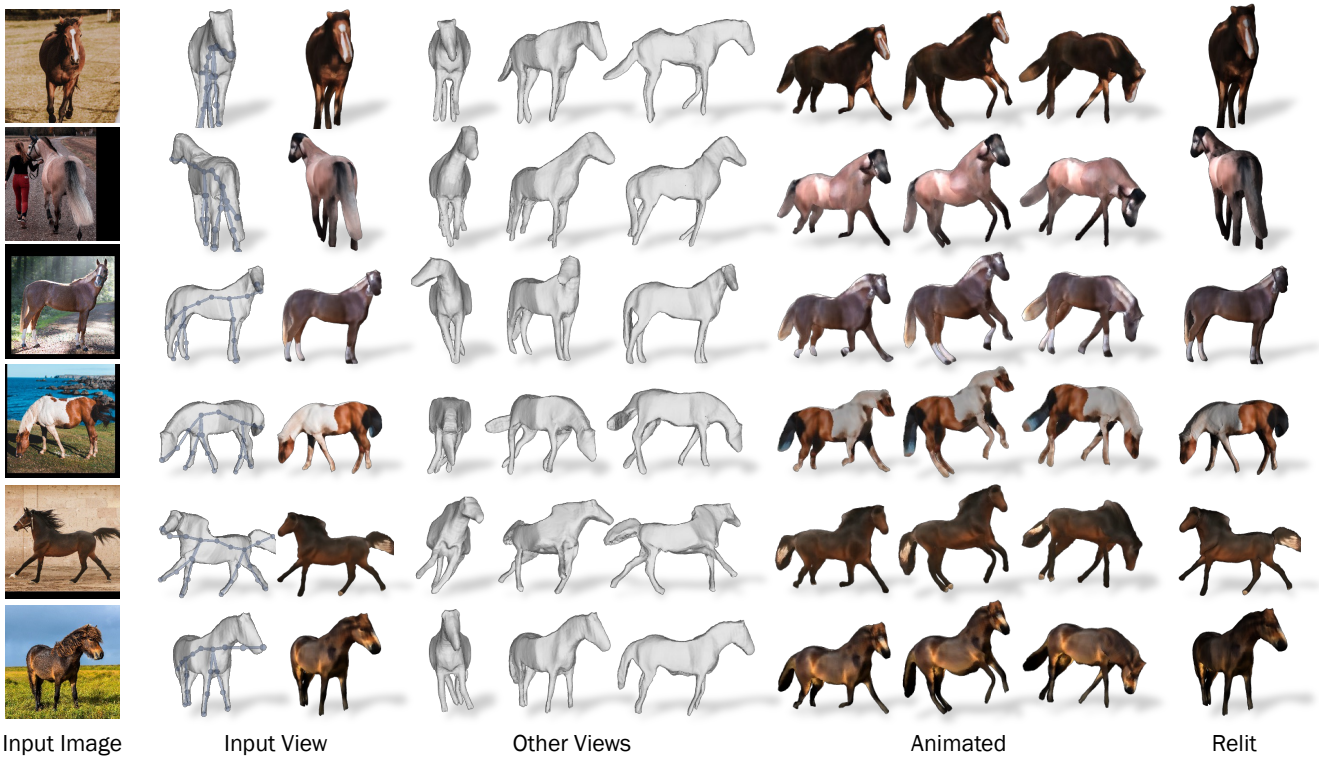

Figure 15. **Reconstruction of Real Horse Images.** We show the predicted mesh from the input view and three additional views. We also demonstrate that our shape can be animated by articulating the estimated skeleton. Finally, as our method decomposes albedo and lightning, our predictions can be easily relit.

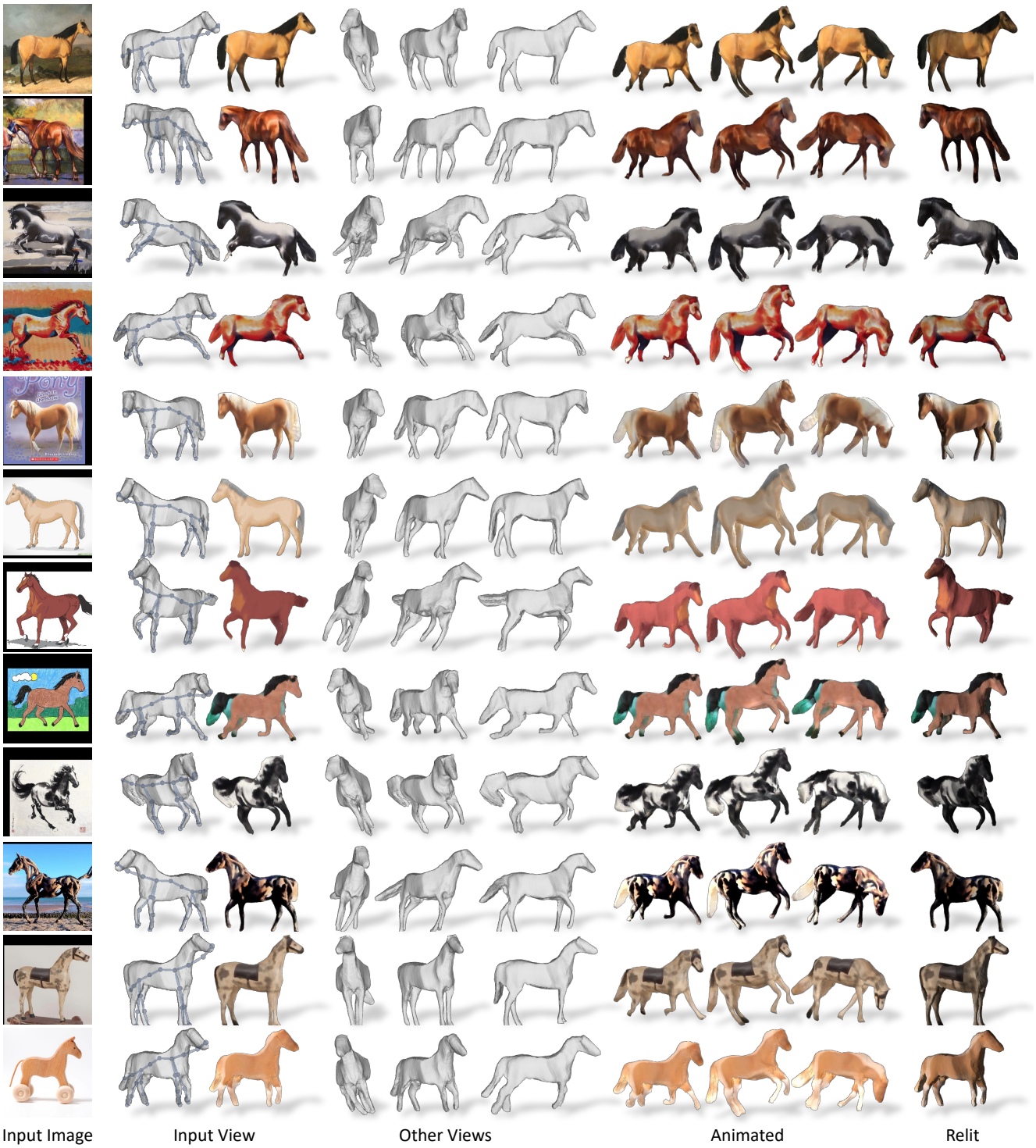

Figure 16. **Reconstruction of Abstract Horse Drawings and Artefacts.** As in Fig. 15, here we show the predicted meshes from the input view and three additional views together with the animated and relit versions. The results demonstrate excellent generalisation of our method on images far from the distribution of the training set which consists only of real horse images.