MagicPony: Learning Articulated 3D Animals in the Wild

Shangzhe Wu*   Ruining Li*   Tomas Jakab*   Christian Rupprecht   Andrea Vedaldi

Visual Geometry Group, University of Oxford

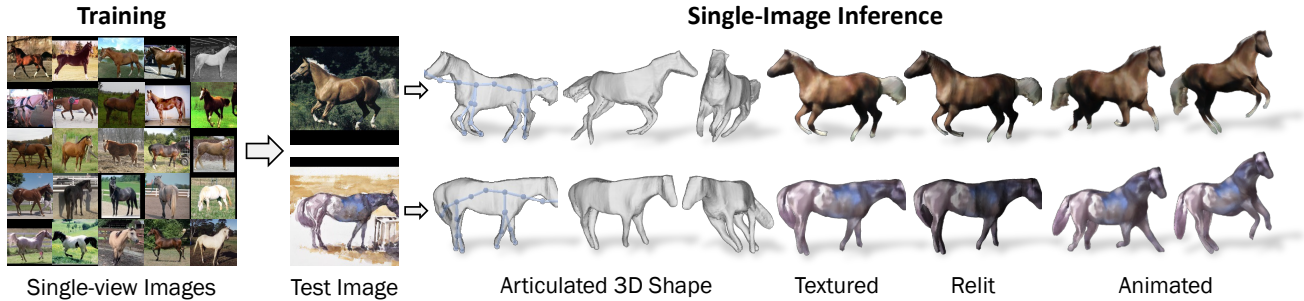{szwu, ruining, tomj, chrisr, vedaldi}@robots.ox.ac.uk

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 the articulated 3D shape of a new instance from a single test image, which can be animated and relit.

Abstract

We consider the problem of predicting the 3D shape, articulation, viewpoint, texture, and lighting of an articulated animal like a horse given a single test image as input. We present a new method, dubbed MagicPony, that learns this predictor 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 local optima in viewpoint estimation, we further introduce a new viewpoint sampling scheme that comes at no additional training cost. MagicPony outperforms prior work on this challenging task and demonstrates excellent generalisation in reconstructing art, despite the fact that it is only trained on real images. The code can be found on the project page at https://3dmagicpony.github.io/.

1. Introduction

Reconstructing the 3D shape of an object from a single image of it requires knowing *a priori* what are the possible shapes and appearances of the object. Learning

such a prior usually requires ad-hoc data acquisition setups [2, 20, 35, 36], involving at least multiple cameras, and often laser scanners, domes and other hardware, not to mention significant manual effort. This is viable for certain types of 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 a 3D prior from 2D images only, which are available in abundance. However, this prior is highly complex and must be learned *while using it to reconstruct in 3D the 2D training data, which is a major challenge.*

In this paper, we propose **MagicPony**, a novel approach to learning 3D models of articulated object categories such as horses and birds with only *single-view* input images for training. We leverage recent progress in unsupervised representation learning, unsupervised image matching, efficient implicit-explicit shape representations and neural rendering, and devise a new auto-encoder architecture that reconstructs the 3D shape, articulation and texture of each object instance from a single image. For training, we only require a 2D segmenter for the object category and a description of the topology and symmetry of its 3D skeleton (*i.e.*, the number and connectivity of bones). We *do not* require apriori knowledge of the objects' 3D shapes, keypoints, viewpoints, or of any other 2D or 3D cue which are often used in prior work [14, 21, 22, 31]. 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 gener-

---

*Equal contribution.

alisation properties, including reconstructing objects in *abstract drawings*, despite being trained on real images only.

In order to learn such disentangled 3D representations simply from raw images, MagicPony addresses a few key challenges. The first challenge is viewpoint estimation. Before the 3D model is available, it is very difficult to assign a viewpoint to the objects in the training images. To tackle this problem, prior works have often assumed additional cues such as 2D keypoint correspondences [21, 22, 31]. Instead, we avoid requiring additional keypoint supervision and implicitly infer noisy correspondences by *fusing* into the 3D model knowledge distilled from DINO-ViT [5], a self-supervised visual transformer network (ViT) [28]. We also develop 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 challenge is how to represent the 3D shape, appearance and deformations of the object. Most prior works have used textured meshes [13, 14, 21, 31, 32, 67, 70], but these are difficult to optimise from scratch, leading to problems that often require ad-hoc heuristics such as remeshing [13, 70]. The other and increasingly popular approach is to use a volumetric representation such as a neural radiance field [3, 40, 52, 63], which can model complex shapes, including manipulating their topology during training. However, this modelling freedom comes at the cost of over-parametrisation, which is particularly problematic in monocular reconstruction and often leads to meaningless short-cut solutions [63]. Furthermore, modelling articulation with a volumetric representation is difficult. A posing transformation is only defined for the object's surface and interior and is more easily expressed from the canonical/pose-free space to the posed space. However, rendering a radiance field requires transforming 3D points off the object surface and in the direction opposite to the posing transformation, which is hard [8].

We address these issues by using a hybrid volumetric-mesh representation based on DMTet [42, 56]. 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.

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; and (3) an efficient multi-hypothesis viewpoint prediction scheme that avoids local optima in reconstruction with no additional cost.

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].

Method	Supervision						Output				
	🕴	📷	🔍	✂	▶▶	🎥	3D	2.5D	Motion	View	Texture
Unsup3D [69]								✓		✓	✓
CSM [30]	✓		✓							✓	
A-CSM [29]	✓		✓						✓	✓	
CMR [21]	(✓)[1]	(✓)[2]	✓	✓			✓			✓	(✓)[3]
U-CMR [14]	✓			✓			✓			✓	✓
UMR[†] [32]				✓			✓			✓	(✓)[3]
ACMR [31]	(✓)[4]	(✓)[2]	✓	✓			✓			✓	(✓)[3]
DOVE [67]			✓	✓	✓	✓			✓	✓	✓
Ours				✓			✓			✓	✓

We compare our method to prior work on several challenging articulated animal categories and show that our method obtains *significantly better* quantitative and qualitative results while using *significantly less* supervision, and also demonstrate 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, 21, 23, 24, 26, 27, 31, 32, 34, 65, 67, 68]. Due to the inherent ambiguity of the problem, these methods usually rely on heavy geometric supervision (see Tab. 1) in addition to object masks, such as 2D keypoints [21, 31], template shapes [14, 29, 30]. UMR [32] 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 [21]. DOVE [67] learns coarse articulated objects by exploiting temporal information from videos with optical flow supervision. All these methods require object masks for supervision, except Unsup3D [69] which exploits bilateral symmetry for reconstructing roughly frontal objects, like faces, and UNICORN [41] which uses a progressive conditioning strategy with a heavily constrained bottleneck, leading to coarse reconstructions. Another emerging paradigm is to leverage generative models [7, 45, 49, 55] that encourage 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, 40, 52, 63] have recently emerged as a powerful volumetric representation for multi-view reconstruction given accurate cameras. A recent line of work, LASR [70], ViSER [71] and BANMo [72], 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 [44] annotations for highly deformable animals. Concurrent work of LASSIE [73] also leverages DINO-ViT [5] image features for supervision but optimises a part-based model on a small collection of images (∼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 [21, 31, 67], which typically requires heavy regularisation, such as As-Rigid-As-Possible (ARAP) [57]. To constrain the space of deformations, many works often utilise lower-dimensional models, such as cages [74] or linear blend skinning controlled with skeletal bones [36, 67] or Gaussian control points [70, 71]. Parametric models, like SMPL [36] and SMAL [76], 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, 46, 48, 50, 51, 58, 59, 72], such as D-NeRF [50], extend NeRF [40] with time-varying components. A-NeRF [58] proposes a radiance field conditioned on an articulated skeleton, while BANMo [72] uses learned Gaussian control points similarly to [70, 71]. 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 [56] 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 on 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 [43] proposes to learn viewpoint distributions implicitly by estimating a probability for each image-pose pair of a fixed object instance. Rel-Pose [75] 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 [25], 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 hierarchical shape representation, from a generic template to articulated instances (Sec. 3.2), and a viewpoint prediction scheme that leverages self-supervised correspondences and efficient multi-hypothesis exploration to avoid local minima (Sec. 3.4), 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) sufficiently *regular* for learning with only weak supervision via image reconstruction.

Most 3D representations are either explicit (triangular meshes [21, 67, 70]) or implicit (volumes, SDFs, radiance fields [19, 39, 40, 47, 66]). Explicit representations are more compact, easily support powerful deformation models like blend skinning, and the geometric smoothness can be easily measured and controlled; however, they are more difficult to optimise and often result in defects like folds during training, which requires occasional re-meshing [13, 70]. Implicit representations are more amenable to gradient-based optimisation, including allowing topological changes, but are more computationally expensive and, due to their high capacity, require substantial supervision, usually in the form of a dense coverage of the different object viewpoints [40, 66] or 3D ground-truth [39, 47].

Here we adopt a representation that combines the advantages of both. The shape of the object is represented implicitly by a neural field that is converted on the fly into an explicit mesh via the *Differentiable Marching Tetrahedra* (DMTet) [9, 42, 56] method. Specifically, the object's shape
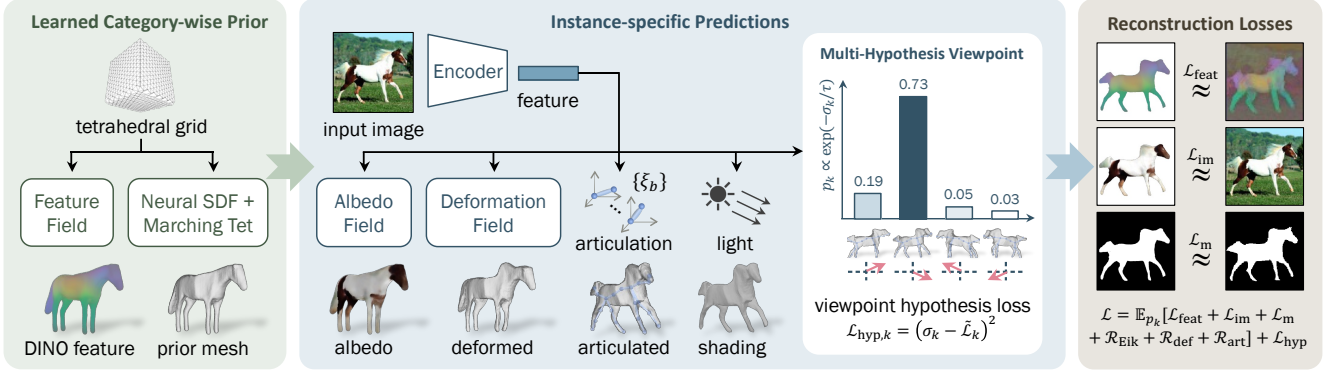
Figure 2. **Training Pipeline.** Given a collection of single-view images of an animal category, our model learns a category-specific prior shape using an implicit-explicit representation, together with a feature field that allows us to fuse self-supervised correspondences through a feature rendering 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.

$S = \{ \boldsymbol{x} \in \mathbb{R}^3 | s(\boldsymbol{x}) = 0 \}$ is defined as the isosurface of the signed distance function (SDF) $s : \mathbb{R}^3 \mapsto \mathbb{R}$, which is implemented by a Multi-Layer Perceptron (MLP) taking 3D coordinates as input. To encourage s to be a signed distance function, we sample random points $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^3$ in space and minimise the Eikonal regulariser [15]:

$$\mathcal{R}_{\text{Eik}}(s) = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x} \in \mathcal{X}} (\|\nabla s(\boldsymbol{x})\|_2 - 1)^2. \quad (1)$$

We then evaluate the function s at the vertices of a tetrahedral grid and extract a watertight mesh from it using DMTet [56], which is fast and differentiable.

We use this implicit-explicit representation to learn a category-wise template shape shared across all instances, and model instance deformations through a hierarchical pipeline, explained next. The function s is initialised with a generic ellipsoid elongated along the z-axis. We denote by $V_{\text{pr}} \subset \mathbb{R}^3$ the vertices of the extracted prior mesh.

3.2. Articulated Shape Modelling and Prediction

The category-wise template V_{pr} is deformed into a specific object instance V_{ins}, and further posed to match the shape V in each input image (Fig. 2). These deformations are predicted from a single image, as discussed below.

Image Encoder. The input image is first encoded by means of DINO-ViT [5], a self-supervised Vision Transformer (ViT) [28]. The ViT extracts from the image a set of key and output tokens, which were found to code, respectively, for correspondences [1] and appearance [61]. Hence, we use the keys to predict the object's deformation and the outputs to predict its appearance. Specifically, we apply the ViT Φ to the input image $I \in \mathbb{R}^{3 \times H \times W}$ and extract the *key* $\Phi_{\text{k}} \in \mathbb{R}^{D \times H_p \times W_p}$ and *output* $\Phi_{\text{o}} \in \mathbb{R}^{D \times H_p \times W_p}$ *tensors* from the last attention layer. These features are local,

with one token per image patch. We use two small convolutional networks to further extract from these a global key $\phi_{\text{k}} = f_{\text{k}}(\Phi_{\text{k}}) \in \mathbb{R}^D$ and output $\phi_{\text{o}} = f_{\text{o}}(\Phi_{\text{o}}) \in \mathbb{R}^D$ vector.

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 $\Delta V_i = f_{\Delta V}(V_{\text{pr},i}, \phi_{\text{k}})$ of each vertex $V_{\text{pr},i}$ is predicted by an MLP $f_{\Delta V}$ from the global image features ϕ_{k}. We exploit the fact that many objects are bilaterally symmetric and enforce both the prior V_{pr} and instance V_{ins} shapes to be symmetric by mirroring the query locations for the MLPs about the yz-plane. We also limit the magnitude of the the deformation V_{ins} by adding the regularizer $\mathcal{R}_{\text{def}}(\Delta V) = \frac{1}{|V_{\text{pr}}|} \sum_i \|\Delta V_i\|^2$.

Posed Shape. In order to match the 3D shape observed in the image, the instance-specific mesh V_{ins} is further deformed by the *posing function* $V_i = g(V_{\text{ins},i}, \xi)$, where ξ denotes the pose parameters. Posing accounts for the deformations caused by the articulation of the object's skeleton; they allow for large but controlled deformations which would otherwise be more difficult to account for by a generic deformation field like ΔV.

We adopt the blend skinning model for posing [6, 37], parameterised by $B - 1$ bone rotations $\xi_b \in SO(3), b = 2, \dots, B$, and the viewpoint $\xi_1 \in SE(3)$. Specifically, we initialise a set of rest-pose joint locations $\mathbf{J}_b$ on the instance mesh using simple heuristics[1]. Each bone b but the root has one parent $\pi(b)$, thus forming a tree structure. Each vertex V_i is associated to the bones by the skinning weights w_{ib}, defined based on relative proximity to each bone (see

---

[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 sup. mat.

sup. mat.for details). The vertices are then posed by the linear blend *skinning equation*:

$$V_i(\xi) = \left(\sum_{b=1}^{B} w_{ib} G_b(\xi) G_b(\xi^*)^{-1} \right) V_{\text{ins},i},$$

$$G_1 = g_1, \ \ G_b = G_{\pi(b)} \circ g_b, \ \ g_b(\xi) = \begin{bmatrix} R_{\xi_b} & \mathbf{J}_b \\ 0 & 1 \end{bmatrix}, \tag{2}$$

where ξ^* denotes the bone rotations at the 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 and ϕ_k. The viewpoint $\xi_1 \in SE(3)$ is predicted separately, as discussed in Sec. 3.4.

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 at rest into the image plane using the predicted viewpoint ξ_1:

$$\boldsymbol{u}_b = \Pi \left[G_{\pi(b)}(\xi_1, \xi_{2:B}^*) \frac{\mathbf{J}_b}{2} \right],$$

and sample the local features from the patch key feature map $\Phi_k(\boldsymbol{u}_b)$ at the projected pixel locations $\boldsymbol{u}_b$.

To predict bone rotations from the image features, we adopt a transformer to account for the inter-dependency of the bones, implied by the underlying skeleton. Specifically, we use a small transformer network [62] f_b that takes as input $B-1$ tokens $\nu_b = (\phi_k, \Phi_k(\boldsymbol{u}_b), b, \mathbf{J}_b, \boldsymbol{u}_b)$ and outputs the bone rotations $(\xi_2, \ldots, \xi_B) = f_b(\nu_2, \ldots, \nu_B)$ parameterised as Euler angles. We also limit the deviation from the rest pose via the regulariser $\mathcal{R}_{\text{art}}(\xi) = \frac{1}{B-1} \sum_{b=2}^{B} \|\xi_b\|^2$.

3.3. Appearance and Shading

We decompose the appearance of the objects into albedo and diffuse shading assuming a Lambertian illumination model. The albedo of the object is modelled in the same "canonical space" as the shape SDF s by a second coordinate MLP $a(\boldsymbol{x}) = f_a(\boldsymbol{x}, \phi_o) \in [0,1]^3$, which takes as input the global output features ϕ_o. The dominant light direction $l \in \mathbb{S}^2$ and ambient and diffuse intensities $k_s, k_d \in \mathbb{R}$ are predicted by a lighting network $f_l(\phi_o)$.

The posed mesh is then rendered into an image via *differentiable deferred mesh rendering*. The renderer first rasterizes all pixels $\boldsymbol{u}$ that overlap with the posed object V, computing the 3D coordinates $\boldsymbol{x}(\boldsymbol{u})$ of the corresponding object point in canonical space; then the MLP f_a is evaluated, once per pixel, to obtain the colour $a(\boldsymbol{x}(\boldsymbol{u}))$. The final pixel colour accounts for shading and is given by

$$\hat{I}(\boldsymbol{u}) = (k_s + k_d \cdot \max\{0, \langle l, n(\boldsymbol{u}) \rangle\}) \cdot a(\boldsymbol{x}(\boldsymbol{u})),$$

where $n(\boldsymbol{u}) \in \mathbb{S}^2$ is the normal of the *posed* mesh V. Using the same differentiable renderer, we also obtain the predicted mask $\hat{M} \in [0,1]^{H \times W}$ of the object.

For training supervision, we use the appearance loss $\mathcal{L}_{\text{im}}(\hat{I}, I, M) = \|\tilde{M} \odot (\hat{I} - I)\|_1$ as well as the mask loss $\mathcal{L}_{\text{m}}(\hat{M}, M) = \|\hat{M} - M\|_2^2 + \lambda_{\text{dt}}(\hat{M} \odot \text{dt}(M))$, where the distance transform $\text{dt}(\cdot)$ is used to obtain better gradients [21, 68], $\tilde{M} = \hat{M} \odot M$ is the intersection of the rendered and ground-truth masks, and λ_{dt} is a weighing factor.

3.4. Viewpoint Prediction

Predicting the object viewpoint is a critical first step for learning 3D shapes but often difficult [14, 21]. Here, we contribute a method to do so robustly and efficiently, which exploits self-supervised correspondences in a multi-hypothesis prediction pipeline, as detailed next.

Fusing Self-Supervised Correspondences. Estimating explicit correspondences without supervision is difficult and brittle; hence, we propose to incorporate correspondence information implicitly and softly. Inspired by N3FF [60], we *fuse* information from the self-supervised ViT features into the 3D model by *learning to render* them.[2] This is done by adding a coordinate network $\psi(\boldsymbol{x}) \in \mathbb{R}^{D'}$ that predicts a D'-dimensional feature vector for each 3D point $\boldsymbol{x}$ in canonical space. This is similar to the albedo network f_a, except that it is shared across all instances of the same object category, and hence there is no dependency on the input image. These canonical features are then projected into a feature image $\hat{\Phi}_k \in \mathbb{R}^{H \times W \times D'}$ using the same differentiable renderer used for the albedo, and trained to minimise a corresponding rendering loss:

$$\mathcal{L}_{\text{feat}}(\hat{\Phi}_k, \Phi_k', \tilde{M}) = \|\tilde{M} \odot (\hat{\Phi}_k - \Phi_k')\|_2^2, \tag{3}$$

where $\Phi_k' = \text{PCA}(\Phi_k)$ is a PCA-reduced (to $D' = 16$ principal components) version of the original DINO-ViT features Φ_k for training efficiency.

Multi-Hypothesis Viewpoint Prediction. Prior works have found that a major challenge in learning the object's 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 only a single one, and thus comes "for free".

We hence task a viewpoint network $f_{\text{vp}}(\phi_k)$ to predict from ϕ_k four viewpoint rotation hypotheses $R_k \in SO(3), k \in \{1, 2, 3, 4\}$, each in one of the four quadrants[3] around the object, as illustrated in Fig. 2. The network also predicts a score σ_k for each hypothesis, used to evaluate the

---

[2]This is also similar to BANMo [72], but we use self-supervised features instead of supervised DensePose [44] ones.

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

probability p_k that hypothesis k is the best of the four options as $p_k \propto \exp(-\sigma_k/\tau)$ using `softmax`, where τ is a temperature parameter, gradually decreased during training.

The naive approach for learning σ_k is to sample multiple hypotheses and compare their reconstruction loss to determine which one is better. However, this is expensive as it requires rendering the model with all hypotheses. Instead, we suggest to *sample* a single hypothesis for each training iteration and simply train σ_k to predict the expected reconstruction loss $\tilde{\mathcal{L}}_k$, by minimising the objective

$$\mathcal{L}_{\text{hyp}}(\sigma_k, \tilde{\mathcal{L}}) = (\sigma_k - \text{sg}[\tilde{\mathcal{L}}_k])^2, \tag{4}$$

where $\text{sg}[\cdot]$ is the stop-gradient operator, and $\tilde{\mathcal{L}}_k$ is the overall reconstruction loss under the sampled k-th hypothesis, given below.

3.5. Training Objective

Given a training sample (I, M), comprising an image and the corresponding object mask, the training objective is computed in two steps. First, the viewpoint hypothesis k is sampled and the reconstruction loss is computed:

$$\tilde{\mathcal{L}}_k = \lambda_{\text{im}}\mathcal{L}_{\text{im}} + \lambda_{\text{m}}\mathcal{L}_{\text{m}} + \lambda_{\text{f}}\mathcal{L}_{\text{feat}}.$$

Then, the regularisation terms and the viewpoint hypothesis loss are added, to obtain the final objective:

$$\mathcal{L} = p_k\tilde{\mathcal{L}}_k + \lambda_{\text{E}}\mathcal{R}_{\text{Eik}} + \lambda_{\text{d}}\mathcal{R}_{\text{def}} + \lambda_{\text{a}}\mathcal{R}_{\text{art}} + \lambda_{\text{h}}\mathcal{L}_{\text{hyp}}.$$

Here $\lambda_{\text{im}}, \lambda_{\text{m}}, \lambda_{\text{f}}, \lambda_{\text{E}}, \lambda_{\text{d}}, \lambda_{\text{a}}, \lambda_{\text{h}}$ are balancing weights.

The gradient of $\mathcal{L}$ with respect to the model parameters is then computed and used to update them. Note that only one viewpoint sample k is taken for each gradient evaluation. To further improve the sampling efficiency, we sample the viewpoint based on the learned probability distribution p_k[4].

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 [67] containing 10.8k 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 [38]. 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 [67] and CUB dataset [64] consisting of 57.9k and 11.7k images respectively. We use an off-the-shelf PointRend [25] 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 subnetworks. 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 [28] 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 f_{vp} identical to f_k, as we find empirically that sharing the encoders 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 hyper-parameters are included in the supplementary material.

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.

Fig. 4 also shows that by finetuning the horse model on other animals without any additional modifications (except disabling articulation for the initial 5k iterations), it gen-

---

[4]We sample viewpoint $k^* = \arg\max_k p_k$ 80% of the time for training efficiency, and uniformly at random 20% of the time.
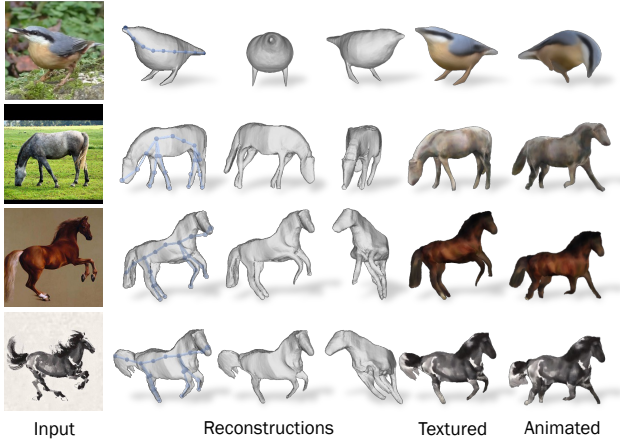
Figure 3. **Single Image Reconstruction.** We show the reconstructed mesh from the input view and two additional views together with predicted texture and animated version of the shape obtained by articulating the estimated skeleton.
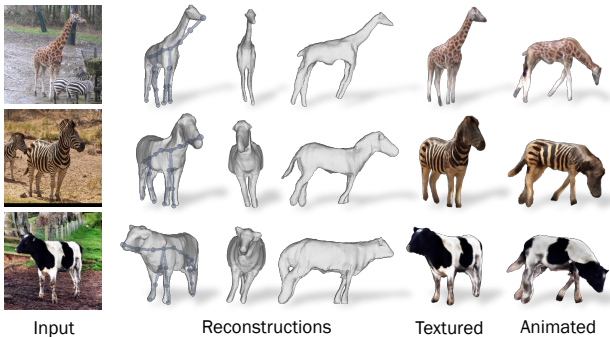

Figure 4. **Reconstructions of Giraffes, Zebras and Cows.** We finetune the horse model on other animal categories, and show that the method generalises well to other animals.

eralises to various animal categories with highly different shapes and appearances, 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 10 seconds. This is also done for giraffes, zebras and cows in Fig. 4, as the training datasets are too small to learn complex appearance. Additional qualitative results are presented in the *supplementary material*.

4.4. Comparison with Prior Work

We compare with previous weakly-supervised methods for 3D reconstruction of deformable objects. The most relevant prior work is UMR [32] and DOVE [67], as they also only require object masks and weak correspondences from either part segmentations (SCOPS) [18] or video training [67]. Our method leverages *self-supervised* DINO-ViT
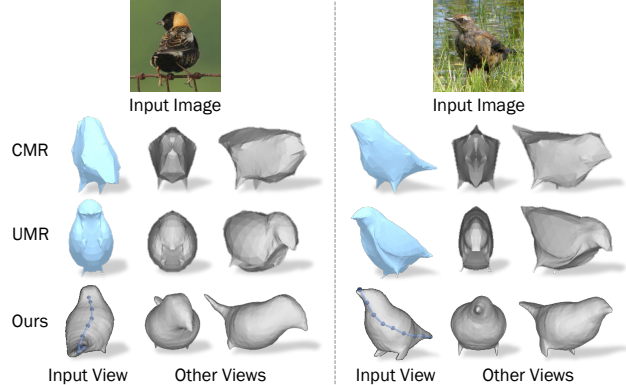

Figure 5. **Comparison on Birds.** Both CMR [21] and UMR [32] often predict inaccurate poses, such as the bird on the left.
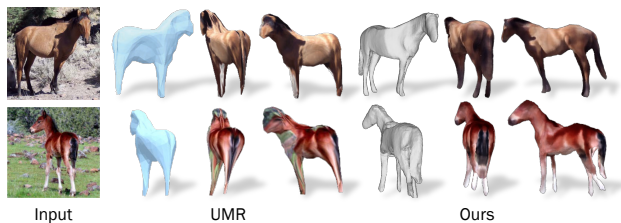

Figure 6. **Comparison with UMR [32] on Horses.** Our model produces much higher quality reconstructions with four legs. UMR is only able to predict symmetrical shapes with two legs.

features, which makes our method the least supervised in this area. We also compare against the well-established baseline of CMR [21], 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 opposed 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 the input image, for a 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 [67], which contains 3D scans of 23 realistic toy bird models paired with 345 photographs of them taken in natural real-world environments. Following [67], we align the predicted

Table 2. **Evaluation on Toy Bird Scans [67].** Baseline results reported by [67]. Annotations: ⍲ template shape, 📷 viewpoint, 🔑 2D keypoint, ✂ mask, ▶▶ optical flow, 🎥 video.

	Supervision	Chamfer Distance (cm) ↓
CMR [21]	📷 🔑 ✂	1.35 $_{\pm 0.81}$
U-CMR [14]	⍲ ✂	1.82 $_{\pm 0.93}$
UMR [32]	✂ + SCOPS	1.24 $_{\pm 0.75}$
DOVE [67]	✂ ▶▶ 🎥	1.51 $_{\pm 0.89}$
Ours	✂ + DINO	**0.79** $_{\pm 0.50}$

Table 3. **Keypoint Transfer on CUB [64] Bird and PASCAL VOC [10] Horse Datasets.** Our model produces superior results on both birds and horses with significantly less supervision.

		PCK@0.1 ↑	
Method	Supervision	Birds	Horses
CMR [21] (from [32])	📷 🔑 ✂	0.473	–
CMR [21]	📷 🔑 ✂	0.546	–
CMR [21]	✂	0.255	–
U-CMR [14]	⍲ ✂	0.359	–
UMR [32]	✂ + SCOPS	0.512	0.244
A-CSM [29]	⍲ ✂	0.426	0.329
DOVE [67]	✂ ▶▶ 🎥	0.447	–
Ours	✂ + DINO	**0.554**	**0.429**

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 to other methods. Our model produces significantly more accurate reconstructions, resulting in a much lower error.

We also evaluate Keypoint Transfer on the CUB [64] and PASCAL VOC [10] benchmarks, a common evaluation metric [21, 29, 32]. We follow the protocol in [29, 32] and sample 20k 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 Correct Keypoints (PCK) metric. As shown in Tab. 3, our model outperforms previous methods, including CMR [21] which requires strong geometric supervision in the form of keypoints.

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. Without either component, the learned viewpoints collapse to only frontal
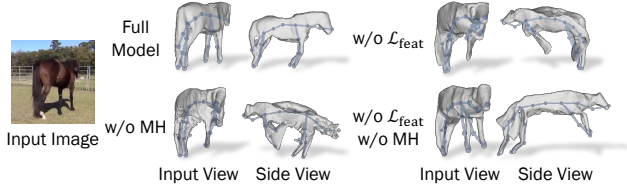

Figure 7. **Ablation Studies.** We train the model without feature loss $\mathcal{L}_{\text{feat}}$, without multi-hypothesis (MH) viewpoint prediction and without both. Without either component, the viewpoint predictions collapse to frontal views, resulting in unnatural shapes.

views, resulting in unnaturally stretched reconstructions. Additional results are provided in the sup. mat.

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 [53, 54]. 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 the sup. mat.

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.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. In *ECCV Workshops*, 2022. 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. 4

[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. 2

[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. 3

[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, 8

[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, 3

[14] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoints without keypoints. In *ECCV*, 2020. 1, 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] Wonbong Jang and Lourdes Agapito. CodeNeRF: Disentangled neural radiance fields for object categories. In *ICCV*, 2021. 3

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

[21] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 1, 2, 3, 5, 7, 8

[22] Abhishek Kar, Shubham Tulsiani, João Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *CVPR*, 2015. 1, 2

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

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

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

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

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

[28] 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

[29] Nilesh Kulkarni, Abhinav Gupta, David F. Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *CVPR*, 2020. 2, 8

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

[31] 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. 1, 2, 3

[32] 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

[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] Nadia Magnenat-Thalmann, E Primeau, and Daniel Thalmann. Abstract muscle action procedures for human face animation. *The Visual Computer*, 1988. 4

[38] 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

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

[40] 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

[41] 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

[42] 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. 2, 3

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

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

[45] 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. 2

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

[47] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 3

[48] 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

[49] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 2

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

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

[52] 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

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

[54] 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

[55] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, 2020. 2

[56] 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

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

[58] 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

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

[60] 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

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

[62] 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

[63] 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

[64] 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

[65] 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

[66] 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. 3

[67] 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

[68] Shangzhe Wu, Ameesh Makadia, Jiajun Wu, Noah Snavely, Richard Tucker, and Angjoo Kanazawa. De-rendering the world's revolutionary artefacts. In *CVPR*, 2021. 2, 5

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

[70] 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

[71] 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

[72] 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

[73] 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

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

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

[76] 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