International Journal of Computer Vision (2023) 131:2623–2634
https://doi.org/10.1007/s11263-023-01819-5

S.I. : COMPUTER VISION APPROACH FOR ANIMAL TRACKING AND MODELING

DOVE: Learning Deformable 3D Objects by Watching Videos

Shangzhe Wu[1] · **Tomas Jakab**[1] · **Christian Rupprecht**[1] · **Andrea Vedaldi**[1]

Received: 29 April 2022 / Accepted: 15 May 2023 / Published online: 15 June 2023

Abstract
Learning deformable 3D objects from 2D images is often an ill-posed problem. Existing methods rely on explicit supervision to establish multi-view correspondences, such as template shape models and keypoint annotations, which restricts their applicability on objects "in the wild". A more natural way of establishing correspondences is by watching videos of objects moving around. In this paper, we present DOVE, a method that learns textured 3D models of deformable object categories from monocular videos available online, without keypoint, viewpoint or template shape supervision. By resolving symmetry-induced pose ambiguities and leveraging temporal correspondences in videos, the model automatically learns to factor out 3D shape, articulated pose and texture from each individual RGB frame, and is ready for single-image inference at test time. In the experiments, we show that existing methods fail to learn sensible 3D shapes without additional keypoint or template supervision, whereas our method produces temporally consistent 3D models, which can be animated and rendered from arbitrary viewpoints. Project page: https://dove3d.github.io/.

Keywords Deformable 3D objects · Unsupervised 3D learning

1 Introduction

In applications, we often need to obtain accurate 3D models of deformable objects from just one or a few pictures of them. This is the case in traditional applications such as robotics, but also, increasingly, in consumer applications, such as content creation for virtual and augmented reality—using everyday pictures and videos taken with a cellphone.

3D reconstruction from a single image, or even a small number of views, is generally very ambiguous and only solvable by leveraging powerful statistical priors of the 3D world. Learning such priors is however very challenging. One approach is to use training data specifically collected for this purpose, for example by using 3D scanners and domes (Choy et al., 2016; Girdhar et al., 2016; Wu et al., 2016; Wang et al., 2018; Groueix et al., 2018; Kato et al., 2018; Mescheder et al., 2019; Park et al., 2019; Saito et al., 2019; Gkioxari et al., 2019) or shape models (Loper et al., 2015; Kanazawa et al., 2018; Zuffi et al., 2017; Sanyal et al., 2019; Zuffi et al., 2019). This is expensive and can be justified only for a few categories such as human bodies and faces that are of particular significance in applications. However, scanning is not a viable approach to cover the huge diversity of objects that exist in the real world.

We thus need to develop methods that can learn 3D deformable objects from as cheap supervision as possible, such as leveraging casually-collected images and videos found on the Internet, or crowdsourced datasets such as CO3D (Reizenstein et al., 2021). Ideally, our system should take as input a collection of such casual images and videos and learn a model capable of reconstructing the 3D shape, appearance and deformation of a new object from a single image of it.

While several authors have looked at this problem before (Choy et al., 2016; Girdhar et al., 2016; Wang et al., 2018; Groueix et al., 2018; Kato et al., 2018; Gkioxari et al., 2019), so far it has always been necessary to make additional simpli-

Shangzhe Wu, Tomas Jakab have contributed equally to this work.

✉ Shangzhe Wu
szwu@robots.ox.ac.uk

✉ Tomas Jakab
tomj@robots.ox.ac.uk

Christian Rupprecht
chrisr@robots.ox.ac.uk

Andrea Vedaldi
vedaldi@robots.ox.ac.uk

[1] Visual Geometry Group, University of Oxford, Oxford, UK

fying assumptions compared to the ideal unsupervised setting described above. These assumptions usually come in the form of additional geometric supervision. The most common one is to require 2D masks for the objects, either obtained manually or via a pre-trained segmentation network such as He et al. (2017), Kirillov et al. (2020). On top of this, there is usually at least one more *additional* form of geometric supervision, such as providing an initial approximate 3D template of the object, 2D keypoint detections, or approximate 3D camera parameters (Kanazawa et al., 2018; Li et al., 2020a; Kokkinos & Kokkinos, 2021a; Kulkarni et al., 2019; Zuffi et al., 2019; Goel et al., 2020; Kokkinos & Kokkinos, 2021b; Niemeyer et al., 2020). There is a small number of works that require no masks or geometric supervision (Wu et al., 2020), but they come with other limitations such as relying on limited viewpoint range.

Our aim in this paper is to learn 3D deformable objects from complex *casual videos* while only using 2D masks and optical flow estimations obtained from off-the-shelf models, removing the additional geometric supervision from expensive manual annotations that are commonly used in prior works (keypoints, viewpoint, and templates). In order to compensate for this lack of geometric information, we propose to learn from casual videos rather than still images, unlike most prior works. While this adds some complexity to the method, using videos has the key advantage to allow one to establish correspondences between different images, for instance by using an off-the-shelf optical flow algorithm. While this information is weaker than externally-provided information such as keypoints, nevertheless it is very helpful in recovering the objects' viewpoint. Note, though, that videos are only used for supervision: our goal is still to learn a model that can reconstruct a new object instance from a single image.

In order to use videos effectively, we make a number of technical contributions. The first one addresses the challenge of estimating the viewpoint of the 3D objects. Prior works addressed this issue by sampling a large number of possible views (Kulkarni et al., 2020; Goel et al., 2020), an approach that (Goel et al., 2020) calls a *camera multiplex*. We find that this is unnecessary. While viewpoint estimation is ambiguous, we show that the ambiguity is mostly restricted to a small space of symmetries induced by the 2D projection of the 3D objects onto the image. The result is that, as the model is learned, only a very small number of alternative viewpoints need to be explored in order to escape form the ambiguity-induced local optima: from, e.g., 40 in Goel et al. (2020) to just two per iteration, which largely reduces memory and time requirements for training.

Our second contribution is the design of the object model. We propose a hierarchical shape representation that explicitly disentangles category-dependent prior shape, instance-dependent deformation, as well as time-dependent articulated and rigid pose. In this way, we automatically factor shape and pose variations at different levels in the video dataset, and leverage instance-specific correspondences within a video and instance-agnostic correspondences across multiple videos. We also enforce a bilateral symmetry on the predicted canonical shape and texture, similar to previous methods (Kanazawa et al., 2018; Goel et al., 2020; Li et al., 2020b; Wu et al., 2020). However, differently from these approaches, which assume symmetry at the level of the object instances, here we assume the canonical (pose free) shapes are symmetric, but individual articulations can be asymmetric (Thewlis et al., 2018; Fernandez-Labrador et al., 2020), which is much more realistic.

We also address the challenge of evaluating these reconstruction methods. Prior works in this area generally lack data with 3D ground truth. Instead, they resort to indirect evaluation by measuring the quality of the 2D correspondences that are established by the 3D models. To address this problem, we create a dataset of views of real-life animal models (toy birds). The data is designed to resemble a subset of the images as found in existing datasets such as CUB (Wah et al., 2011); however, it additionally comes with 3D scans of the objects, which can be used to test the quality of the 3D reconstructions directly. We use this data to evaluate our and

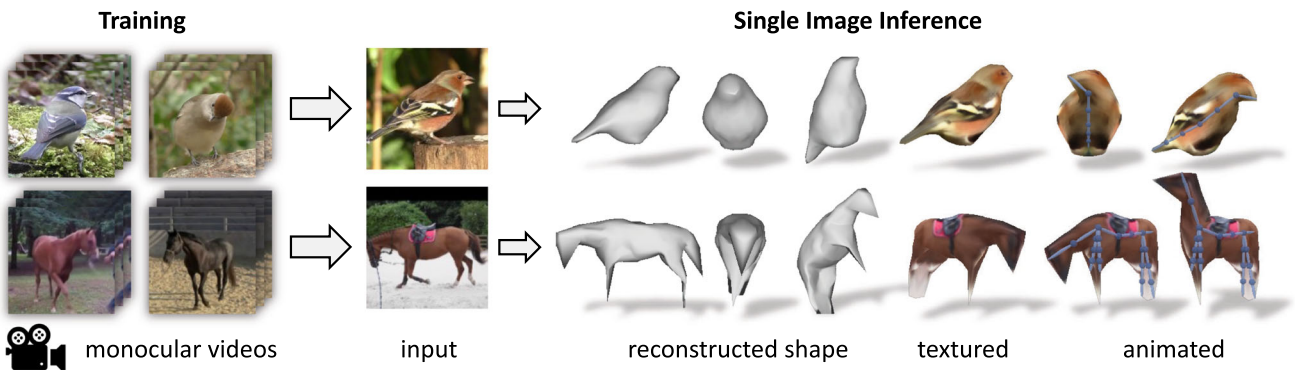

Fig. 1 DOVE—deformable objects from videos. Given a collection of video clips of an object category as training data, we learn a model that is able to predict a textured, articulated 3D mesh of the object from a single input image

several state-of-the-art algorithms without the need for proxy metrics such as keypoint re-projection error that are insufficient to assess the quality of a 3D reconstruction. We hope that this data will be useful for future work in this area.

Overall, our method can successfully learn good 3D shape predictors from videos of animals such as birds and horses. Compared to prior work, our method produces better 3D shape reconstructions, as measured on the new benchmark, when not using additional geometric supervision.

2 Related Work

We divide the vast literature of related work into two parts. The first one focuses on learning based approaches for 3D reconstruction with limited supervision. The second parts highlights related work for 3D reconstruction from images and video.

2.1 Unsupervised and Weakly-Supervised 3D Reconstruction

A primary motivation for introducing machine learning in 3D reconstruction is to enable reconstruction from single views, which necessitates learning suitable shape priors. In particular, we focus the discussion on unsupervised and weakly-supervised methods that do not require explicit 3D ground-truth for training. Early unsupervised work include monocular depth predictors trained from egocentric videos of rigid scenes (Garg et al., 2016; Zhou et al., 2017).

Others have explored weakly-supervised methods for learning full 3D meshes of object categories (Kato et al., 2018; Kanazawa et al., 2018; Liu et al., 2019; Kato & Harada, 2019; Wang et al., 2018; Henderson & Ferrari, 2019; Goel et al., 2020; Li et al., 2020b, a; Wu et al., 2021; Kokkinos & Kokkinos, 2021b, a). Many of these methods learn from still images and generally require masks and other additional supervision or prior assumptions, summarized in Table 1. In particular, **CMR** (Kanazawa et al., 2018) uses 2D keypoint annotations (in addition to masks) to initialize shape and viewpoint using Structure-from-Motion (SfM). This is extended in the follow-up works in various ways. **U-CMR** (Goel et al., 2020), **TTP** (Kokkinos & Kokkinos, 2021b) and **IMR** (Tulsiani et al., 2020) replace the keypoint annotations with a category-specific template shape beforehand. With the template shape, extensive viewpoint sampling (camera multiplex) can be done to search for the best camera viewpoint for each training image (Goel et al., 2020). **UMR** (Li et al., 2020b) instead uses part segmentations from SCOPS (Hung et al., 2019), which also requires supervised ImageNet pretraining. **VMR** (Li et al., 2020a) extends CMR with asymmetric deformation, and introduces a test-time adaptation procedure on individual videos by enforcing temporal consistency on the predictions produced by a pre-trained CMR model. Note that we use videos to learn a 3D shape model *from scratch*, whereas VMR starts with a pre-trained model and only performs online adaptation on videos. **CSM** (Kulkarni et al., 2019) and **articulated CSM** (Kulkarni et al., 2020) learn to pose an externally-provided (articulated) 3D template of an object category to images. Unsup3D (Wu et al., 2020) learns symmetric objects, like faces, without masks, but only with limited viewpoint variation.

Adversarial learning has also been explored to replace the need of multi-views for training (Kudo et al., 2018; Chen et al., 2019; Henzler et al., 2019; Nguyen-Phuoc et al., 2019, 2020; Ye et al., 2021; Schwarz et al., 2020; Niemeyer & Geiger, 2021; Zhang et al., 2021; Chan et al., 2021; Pan et al., 2021). The idea is to use a discriminator network to tell whether or not arbitrarily generated views of the learned 3D model are plausible, which provides signals to learn the geometry. Although this approach does not require viewpoint annotations for individual images, it does rely on a reasonable approximation of the viewpoint distribution in the training data, from which random views are generated. Overall, promising results can be achieved on synthetic data as well as a few real object categories, but general methods usually recover only coarse 3D shapes or 3D feature volumes that are difficult to extract.

2.2 Reconstruction from Multiple Views and Videos

Most works using multiple views and videos focus on reconstructing individual instances of an object. Classic SfM methods (Faugeras & Luong, 2001; Hartley & Zisserman, 2004) use multiple views of a rigid scene, with pipelines such as KinectFusion (Newcombe et al., 2011) and DynamicFusion (Newcombe et al., 2015) integrating depth sensors for reconstructing dense static and deformable surfaces. Neural implicit surface representations have recently emerged for multi-view reconstruction (Yariv et al., 2020; Wang et al., 2021; Oechsle et al., 2021). NeRF (Mildenhall et al., 2020) and its deformable extensions (Park et al., 2021; Gafni et al., 2021; Tretschk et al., 2021; Raj et al., 2021; Noguchi et al., 2021; Pumarola et al., 2020) synthesize novel views from densely sampled multi-views of a static or mildly dynamic scene using a Neural Radiance Field, from which explicit coarse 3D geometry can be further extracted. A more recent line of work, such as LASR (Yang et al., 2021a) and ViSER (Yang et al., 2021b), optimizes a single 3D deformable model on an individual video sequence, using mask and optical flow supervision. BANMo (Yang et al., 2022) further extends the pipeline and optimizes over a few video sequences of the same object instance, with the help of a pretrained DensePose (Neverova et al., 2020) model. However, these optimization-based models are typically sensitive to the quality of the sequences and tends to fail when only limited views are

Table 1 Related work overview

Method	Supervision						Output				
	template	viewpoint	2D keypoint	object mask	optical flow	video	3D	2.5D	Motion	Viewpoint	Texture
VMR* (Li et al., 2020a)	(✓)[1]			✓		✓	✓		✓	✓	(✓)[2]
LASR* (Yang et al., 2021a)				✓	✓	✓	✓		✓	✓	✓
ViSER* (Yang et al., 2021b)				✓	✓	✓	✓		✓	✓	✓
BANMo* (Yang et al., 2022)		(✓)[3]	(✓)[3]	✓	✓	✓	✓		✓	✓	✓
Unsup3D (Wu et al., 2020)								✓		✓	✓
CSM (Kulkarni et al., 2019)	✓			✓						✓	
CMR (Kanazawa et al., 2018)	(✓)[4]	(✓)[4]	✓	✓			✓			✓	(✓)[2]
U-CMR (Goel et al., 2020)	✓			✓			✓			✓	✓
IMR (Tulsiani et al., 2020)	✓			✓			✓			✓	(✓)[2]
TTP (Kokkinos & Kokkinos, 2021b)	✓			✓			✓			✓	(✓)[2]
UMR† (Li et al., 2020b)				✓			✓			✓	(✓)[2]
VMR (Li et al., 2020a)	(✓)[1]	(✓)[4]	✓	✓			✓		✓	✓	(✓)[2]
Ours				✓	✓	✓	✓		✓	✓	✓

Annotations: template shape, viewpoint, 2D keypoint, object mask, optical flow, video, *optimizes a single object instance over a single or a few sequences, [1]shape bases initialized from CMR (Kanazawa et al., 2018), [2]outputs texture flow, [3]obtained from DensePose (Neverova et al., 2020), [4]obtained from keypoints using SfM, †UMR (Li et al., 2020b) relies on part segmentations from SCOPS (Hung et al., 2019)

observed (see Fig. 6). In contrast, by learning priors over a video dataset, our model can perform inference on a single image.

Other works that learn 3D categories form videos typically require some shape prior, such as a parametric shape model (Loper et al., 2015; Paysan et al., 2009), and hence mostly focus on reconstruction of human bodies or faces (Tung et al., 2017; Arnab et al., 2019; Doersch & Zisserman, 2019; Kanazawa et al., 2019; Zhang et al., 2019; Feng et al., 2018; Tran & Liu, 2019; Kokkinos & Kokkinos, 2021a; Zuffi et al., 2019). Novotný et al. (2017) and Henzler et al. (2021) consider turn-table like videos to learn to reconstruct rigid object categories. In contrast, our method learns a 3D shape model of a deformable object category *from scratch* using videos.

3 Method

Our goal is to learn a function $(V, \xi, T) = f(I)$ that, given a *single image* $I \in \mathbb{R}^{3\times H\times W}$ of an object, predicts its 3D shape V (a mesh), its pose ξ (either a rigid transformation or full articulation parameters) and its texture T (an image). We describe below the key ideas in our method and refer the reader to the sup. mat. for details.

While the predictor f is monocular, we supervise it by making use of video sequences $\mathcal{I} = \{I_t\}_{t=1,\dots,|\mathcal{I}|}$, where t denotes time. For this, we use a *photo-geometric auto-encoding approach*. Let $M \in \{0,1\}^{H\times W}$ be the 2D mask of the object in image I, which we assume to be given. The model $(V, \xi, T) = f(I)$ encodes the image as a set of photo-geometric parameters; from these, an handcrafted rendering function $(\hat{I}, \hat{M}) = \mathcal{R}(V, \xi, T)$ reconstructs the image $\hat{I}$ and the mask $\hat{M}$. For supervision, the rendered image and the rendered mask is compared to the given ones via two losses:

$$L_{\text{im}} = \lambda_{\text{im}} \|\hat{M} \odot (\hat{I} - I)\|_1, \tag{1}$$

$$L_{\text{mask}} = \lambda_{\text{mask}} \|\hat{M} - M\|_2^2, \tag{2}$$

where λ_{im} and λ_{mask} weigh each loss. Note that the image loss is restricted to the predicted region as the model only represents the object but not the background. Figure 2 gives an overview of the training pipeline.

3.1 Solving the Viewpoint Ambiguity

Decomposing a single image into shape, pose and appearance is ambiguous, which is a major challenge that can easily result in poor reconstructions. Some prior works have addressed this issue by sampling a large number of viewpoints during training, thus giving the optimizer a chance to avoid local optima. However, this is a slow process that requires testing a large number of hypotheses at every iteration (*e.g.* 40 in Goel et al. (2020)) and requires a precise template shape to understand the differences between small viewpoint changes.

Here we note that this is likely unnecessary. The key observation is that the ambiguities arising from image-based reconstruction are not arbitrary; instead, they tend to concentrate around specific symmetries induced by the projection of a 3D object onto an image.

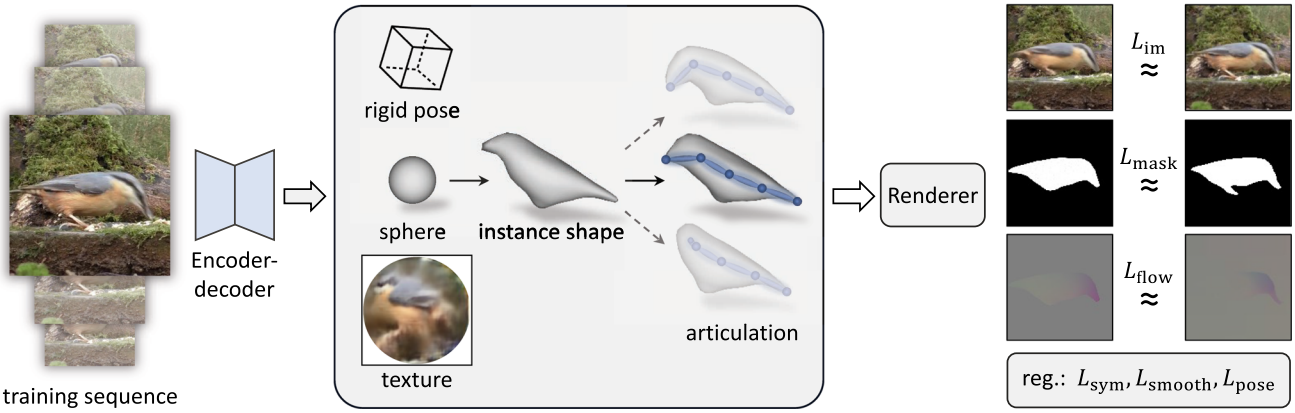

Fig. 2 Training pipeline. From a single frame of a video, we predict the 3D pose, shape and texture of the object. The shape is further disentangled into category shape, instance shape and deformation using linear blend skinning. Using a differentiable rendering step, we can train the model end-to-end by reconstructing the image and by enforcing temporal consistencies

The image to the right illustrates this idea. Here, given only the mask M, one is unable to choose between the object pose ξ or its mirrored variant $q\xi$, where q is a suitable 'mirror mapping' that rotates the pose back to front (see sup. mat. for details). We argue that, before developing a more nuanced understanding of appearance, the model f is similarly undecided about the pose of the 3D object; however, the number of choices is very limited: either the current prediction $\xi = f_\xi(I)$ is correct, or its mirrored version $q\xi$ is.

Concretely, during training we evaluate the loss $L(V, \xi, T)$ for the model prediction and the loss $L(V, q\xi, T)$ for the mirrored pose. We find the better of the two poses $\xi^* = \arg\min_{\hat{\xi}\in\{\xi, q\xi\}} L(V, \hat{\xi}, T)$ and optimize the loss:

$$L_{\text{pose}} = \lambda_{\text{pose}} \|\xi - \xi^*\|_2^2 . \tag{3}$$

In this way, the model is encouraged to flip its prediction when $L(V, q\xi, T) < L(V, \xi, T)$. This assures that the model eventually learns the correct pose and does not rely on the flipping towards the end of the training.

3.2 Learning from Videos

We exploit the information in videos by noting that the shape V and texture T of an object are invariant over time, with any time-dependent change limited to the pose ξ. Hence, given a sequence of images $\mathcal{I} = \{I_t\}_{t=0,\dots,|\mathcal{I}|}$ of the same object and corresponding frame-based predictions $(V_t, \xi_t, T_t) = \Phi(I_t)$, we feed the rendering function $(\hat{I}_t, \hat{M}_t) = R(\bar{V}, \xi_t, \bar{T})$ with the shape and texture averages $\bar{V} = \frac{1}{|\mathcal{I}|}\sum_{t=1}^{|\mathcal{I}|} V_t$ and $\bar{T} = \frac{1}{|\mathcal{I}|}\sum_{t=1}^{|\mathcal{I}|} T_t$. The idea is that, unless shape and texture agree across predictions, their averages would be blurry and result in poor renderings. Hence, minimizing the rendering loss indirectly encourages these quantities to be consistent over time.

Furthermore, while the pose ξ_t does vary over time, pose changes must be compatible with image-level correspondences. Specifically, let $F_t \in \mathbb{R}^{H\times W\times 2}$ be the optical flow measured between frames I_t and I_{t+1} by an off-the-shelf method such as RAFT (Teed & Deng, 2020). We can render the flow $\hat{F}_t = \mathcal{R}(V, \xi_t, \xi_{t+1})$ by computing the displacement of the object vertices V as a pose change from ξ_t to ξ_{t+1}. We can then add the flow reconstruction loss

$$L_{\text{flow}}(\hat{F}_t, F_t) = \lambda_{\text{flow}} \|M_t \odot (\hat{F}_t - F_t)\|_2^2 , \tag{4}$$

to encourage consistent motion of the object. Its influence is controlled by the weight λ_{flow}.

3.3 Hierarchical Shape Model

Next, we flesh out the shape model. The shape $V \in \mathbb{R}^{3\times K}$ is given by K mesh vertices and represents the shape of a specific object instance in a *canonical* pose. It is obtained by the predictor $f_V(I) = V_{\text{base}} + \Delta V_{\text{tmpl}} + \Delta V(I)$ where: V_{base} is an initial fixed shape (a sphere), ΔV_{tmpl} is a learnable matrix (initialized as zeros) such that $V_{\text{tmpl}} = V_{\text{base}} + \Delta V_{\text{tmpl}}$ gives an average shape for the category (template), and $\Delta V(I)$ is a neural network further deforming the this template into the specific shape of the object seen in image I. We further restrict V, which is the rest pose, to be bilaterally symmetric by only predicting half of the vertices and obtaining the remaining half via mirroring along the x axis. Note that, while in many prior works the category-level template V_{tmpl}

is given to the algorithm, here this is learned automatically from a sphere.

Finally, the shape V is transformed into the actual mesh observed in the image by a *posing function* $g(V, \xi)$. We work with two kinds of such functions. The first one is a simple rigid motion $g(V, \xi) = g_\xi V$, $g_\xi \in SE(3)$. This is used in an initial warm-up phase for the model to allow it to learn a first version of the template V automatically.

In a second learning phase, we further enrich the model to capture complex articulations of the shape. There are a number of possible parameterizations that could be used for this purpose. For instance, Kokkinos and Kokkinos (2021a) automatically initializes a set of keypoints via spectral analysis of the mesh. Here, we initialize instead a traditional skinning model, given by a system of bones $b \in \{1, \dots, B\}$, ensuring inelastic deformations.

The skinning model is specified by: the bone topology (a tree), the joint location $\mathbf{J}_b \in \mathbb{R}^3$ of each bone with respect to the parent bone, the relative rotation $\xi_b \in SO(3)$ of that bone with respect to the parent, and a row-stochastic matrix of weights $w \in [0, 1]^{K \times B}$ specifying the strength of association of each mesh vertex to each bone. Of these, only the topology is chosen manually (e.g, to account for a different number of legs for objects in the category). The joint locations $\mathbf{J}_b$ and the skinning weights w are set automatically based on a simple heuristic (described in sup. mat.).

While topology, $\mathbf{J}_b$ and w are fixed, the joint rotation $\xi_b \in SO(3), b = 2, \dots, B$ and the rigid pose $\xi_1 \in SE(3)$ are output by the predictor f to express the deformation of the object as it changes from image to image.

3.4 Appearance Model and Rendering

We model the appearance of the object using a texture map $T \in \mathbb{R}^{3 \times H_T \times W_T}$. The vertices of the base mesh V_{base} are assigned to fixed texture uv-coordinates and the texture inherits the symmetry of the base mesh. Given the posed mesh $g(V, \xi)$ and the texture T, we *render* an image $(\hat{I}, \hat{M}) = \mathcal{R}(V, \xi, T)$ of the object using standard perspective-correct texture mapping with barycentric coordinates using the PyTorch3D differentiable mesh renderer (Ravi et al., 2020).

3.5 Symmetry and Geometric Regularizers

An important property of object categories is that they are often symmetric. This does not mean that individual object instances are symmetric, but that the space of objects is Thewlis et al. (2018). In other words, if image I contains a valid object, so does the mirrored image mI. Furthermore, given the photo-geometric parameters $(V, \xi, T) = f(I)$ for I, the parameters for mI must be given by $f(mI) = (mV, m\xi, Tm)$ where $mV = V$ (because the rest shape is assumed symmetric), Tm is the flipped texture image and $m\xi$ is a mirrored version of the pose. Hence, we additionally enforce the pose predictor to satisfy this structure by minimizing the loss $L_{\text{sym}} = \lambda_{\text{sym}} \| f_\xi(mI) - m(f_\xi(I)) \|_2^2$, weighted by λ_{sym}.

Note the relationship between the mirroring operators $q\xi$ in Sect. 3.1 and $m\xi$ here: they are the same, up to a further rigid body rotation. The effect is that q appears to rotate the object back to front, and m left to right. This is developed formally in the sup. mat.

We further regularise learning the mesh V via a loss $L_{\text{smooth}}(V, V_{\text{tmpl}})$ which includes: the ARAP loss (Sorkine & Alexa, 2007) between V and the template V_{tmlp}, ensuring that they do not diverge too much, and a Laplacian and mesh normal smoothers for V.

3.6 Learning Formulation

Given a video $\mathcal{I} = \{I_t\}_{t=1,\dots,|\mathcal{I}|}$, the overall learning loss is thus:

$$L = L_{\text{im}} + L_{\text{mask}} + L_{\text{flow}} + L_{\text{sym}} + L_{\text{smooth}} + L_{\text{pose}}.$$

In practice, we found it important to warm up the model, activating increasingly more refined model components as training progresses. This can be seen as a sort of coarse-to-fine training strategy.

Learning thus uses the following schedule in three phases: (1) *Shape learning*: the basic model with no instance-specific deformation (i.e., $V = V_{\text{tmpl}}$), no bone articulation and only the mask loss is optimized. This is to allow the model to learn an initial template shape V_{tmpl} of the object category and roughly register the viewpoints of the training images to a canonical frame. (2) *Pose Rectification*: the pose rectification loss L_{pose} is then activated to correct the wrong viewpoint predictions due to the front-to-back ambiguity in Sect. 3.1. (3) *Full model*: finally, the bones are instantiated, and the instance deformation, skinning models and appearance loss are also activated in order to learn the full articulated model.

4 Experiments

We perform an extensive set of experiments to evaluate our method and compare to the state of the art. To this end, we collect three datasets, two of real animals, birds and horses, and one using toy birds of which we can obtain ground truth 3D scans.

4.1 Dataset and Implementation Details

Video Datasets

We experiment with two types of objects: birds and horses. For each category, we extract a collection of short video clips

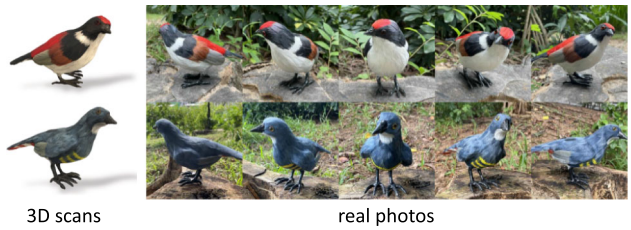

Fig. 3 Examples of the 3D Toy bird dataset. Each bird toy was 3D scanned and the photographed "*in the wild*"

from YouTube. The exact links to these videos and the preprocessing details are included in the sup. mat. We use the off-the-shelf PointRend model (Kirillov et al., 2020) to detect and segment the object instances, remove the frames where the object is static, and automatically split the remaining frames into short clips, each containing one single object. The frames and the masks are then cropped around the objects and resized to 128×128 for training. We also run the off-the-shelf RAFT model (Teed & Deng, 2020) on the full frames to estimate optical flow between consecutive frames, and account for the cropping and resizing to obtain the correct optical flow for the crops. This procedure creates $1,962$ and 114 short clips of birds and horses respectively, each containing 16 to a few hundred frames with paired image, mask and flow. We randomly split them into $1,767/195$ and 103/11 training/testing sequences for birds and horses respectively.

3D Toy Bird Dataset

In order to properly evaluate and compare the quality of the reconstructed 3D shapes produced by different methods, we introduce a 3D Toy Bird Dataset, which consists of ground-truth 3D scans of realistic toy bird models and photographs of them taken in real world environments. Figure 3 shows examples of the dataset. Specifically, we obtain 23 toy bird models, and used Apple RealityKit Object Capture API [82] to capture accurate 3D scans from turn-table videos. For each model, we then take 5 photographs from different viewpoints in 3 different outdoor scenes, resulting in a total of 345 images. We will release the dataset and ground-truth for future benchmarking.

Implementation Details

Our reconstruction model is implemented using three neural networks (f_V, f_ξ, f_T) as well as a set of of trainable parameters for the categorical prior shape ΔV_{tmpl}. The shape network f_V and the rigid pose network f_ξ are simple encoders with downsampling convolutional layers that take in an image and predict vertex deformations ΔV_{ins}, skinning parameters $\xi_{2:B}$, and rigid pose ξ_1 and $\mathbf{J}_1$ as flattened vectors. The texture network f_T is an encoder-decoder that predicts the texture map T from an image. We use Adam optimizers with a learning rate of 10^{-4} for all networks, and a learning rate 0.01 for the category shape parameters ΔV_{tmpl}. We use a symmetric ico-sphere as the initial mesh. For each training iteration, we randomly sample 8 consecutive frames from 8 sequences. The models are trained in three phases described in Sect. 3.6. All details are included in the sup. mat.

4.2 Qualitative Results

Figure 4 shows qualitative 3D reconstruction results obtained from our model. Note that videos are no longer needed during inference and the shown predictions come from a single frame. Despite not requiring any explicit 3D, viewpoint or keypoint supervision, our model learns to reconstruct accurate 3D shapes from only monocular training videos. The reconstructed 3D meshes can be animated with our skinning model by transforming the bones of the learned shape. This animation can also be transferred between instances.

4.3 Comparisons with State-of-the-Art Methods

We compare our model with a number of state-of-the-art learning-based reconstruction methods, including CMR (Kanazawa et al., 2018), U-CMR (Goel et al., 2020), UMR (Li et al., 2020b) and VMR (Li et al., 2020a). CMR requires 2D keypoint annotations for initializing the 3D shape and viewpoints and also for the training loss. U-CMR removes keypoint supervision but requires a 3D template shape, and UMR replaces that with part segmentation maps from

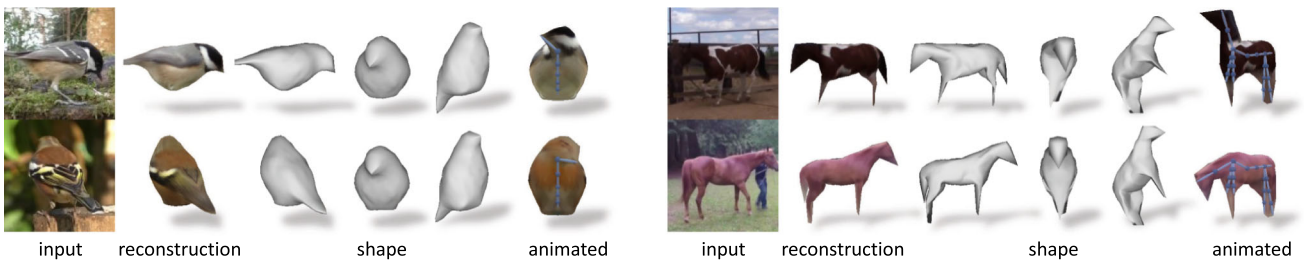

Fig. 4 Qualitative examples. We show multiple views of the reconstructed mesh together with a textured view and animated version of the bird that we obtained by rotating the learned bones. We find that the model is able to recover the shape well even when seen from novel viewpoints. The animation is able to generate believable poses

SCOPS (Hung et al., 2019) which relies on supervised ImageNet pretraining. VMR (Li et al., 2020a) allows for deformations but it requires the same level of supervision as CMR. All of them rely on external geometric supervision to establish correspondences for learning 3D shapes. We train all these methods on our video dataset with only mask supervision and show that without the additional supervision, all these methods reconstruct poor shapes. We also finetune their models pre-trained on CUB (Wah et al., 2011) with the required keypoint, camera view or template shape supervision on our bird video dataset. Finally, we also train UMR from scratch on our bird video dataset with SCOPS predictions obtained from the pre-trained SCOPS model.

On 3D Toy Bird Scans

Our toy scan dataset allows for a direct evaluation of shape prediction. We first scale the predicted shapes to match the volume of the scans and roughly align the canonical pose of each method to the scans manually. Each individual predicted shape is further aligned to the ground-truth scan using Iterative Closest Point (ICP) (Besl & McKay, 1992) and the symmetric (average of scan-to-object and object-to-scan) Chamfer distance is reported in centimeters, Table 2, by assuming the width of each bird to be 10 cm. While the reconstruction quality of other methods is good when trained with more geometric supervision, it degrades strongly without this training signal resulting in worse reconstructions when compared to our method. Note that this metric evaluates the individual shape predictions regardless the viewpoints. Next we evaluate the consistency across views.

On Bird Video Dataset

Since we do not have ground-truth 3D shape and viewpoints for direct evaluation on our video test set, we measure reconstruction quality via a mask *forward* projection accuracy from one frame to another, using the object masks predicted

Table 2 Evaluation on toy bird scans

	Supervision	Chamfer Distance (cm) ↓
CMR (Kanazawa et al., 2018) (finetuned)	📷 🔑 ✂	1.35 ±0.81
U-CMR (Goel et al., 2020) (finetuned)	🧍 ✂	1.82 ±0.93
VMR (Li et al., 2020a) (finetuned)	🧍 📷 🔑 ✂	1.28 ±0.69
UMR (Li et al., 2020b) (finetuned)	✂+SCOPS	1.24 ±0.75
CMR (Kanazawa et al., 2018)	✂	5.94 ±10.33
U-CMR (Goel et al., 2020)	✂	4.36 ±1.56
VMR (Li et al., 2020a)	✂	1.90 ±0.96
UMR (Li et al., 2020b)	✂	2.26 ±1.12
UMR (Li et al., 2020b)	✂+SCOPS	1.82 ±0.93
Ours	✂ ⏩ 🎥	**1.51** ±0.89

The bold number indicates the best results

Shape reconstruction quality measured by bi-directional Chamfer Distance between predicted shape and ground-truth scans. The lower the better. 🧍 template shape, 📷 viewpoint, 🔑 2D keypoint, ✂ mask, ⏩ optical flow, 🎥 video. "finetuned" indicates pretrained models finetuned on our video dataset

Table 3 Mask forward projection IoU

Frame offset	Supervision	$\Delta t = 0$	$\Delta t = 5$	$\Delta t = 20$
CMR (Kanazawa et al., 2018) (finetuned)	📷 🔑 ✂	0.770 ±0.13	0.722 ±0.13	0.712 ±0.13
U-CMR (Goel et al., 2020) (finetuned)	🧍 ✂	0.790 ±0.06	0.761 ±0.07	0.758 ±0.07
VMR (Li et al., 2020a) (finetuned)	🧍 📷 🔑 ✂	0.807 ±0.08	0.752 ±0.08	0.737 ±0.09
UMR (Li et al., 2020b) (finetuned)	✂+ SCOPS	0.847 ±0.05	0.782 ±0.07	0.772 ±0.07
CMR (Kanazawa et al., 2018)	✂	0.634 ±0.06	0.605 ±0.11	0.596 ±0.11
U-CMR (Goel et al., 2020)	✂	0.725 ±0.05	0.714 ±0.06	0.700 ±0.06
VMR (Li et al., 2020a)	✂	0.777 ±0.05	0.720 ±0.07	0.700 ±0.09
UMR (Li et al., 2020b)	✂	0.853 ±0.04	0.798 ±0.06	0.788 ±0.06
UMR (Li et al., 2020b)	✂+ SCOPS	0.830 ±0.04	0.766 ±0.07	0.753 ±0.07
Ours (articulation fixed)	✂ ⏩ 🎥	**0.855** ±0.07	0.810 ±0.07	0.805 ±0.07
Ours (articulation transferred)	✂ ⏩ 🎥	**0.855** ±0.07	**0.844** ±0.07	**0.845** ±0.07

The bold numbers indicate the best results

Shape reconstruction quality and temporal consistency measured by projecting the shape predicted at frame t to a different pose at a *future* frame $t + \Delta t$ and comparing the masks at $t + \Delta t$. The higher the better. "finetuned" indicates pretrained models finetuned on our video dataset

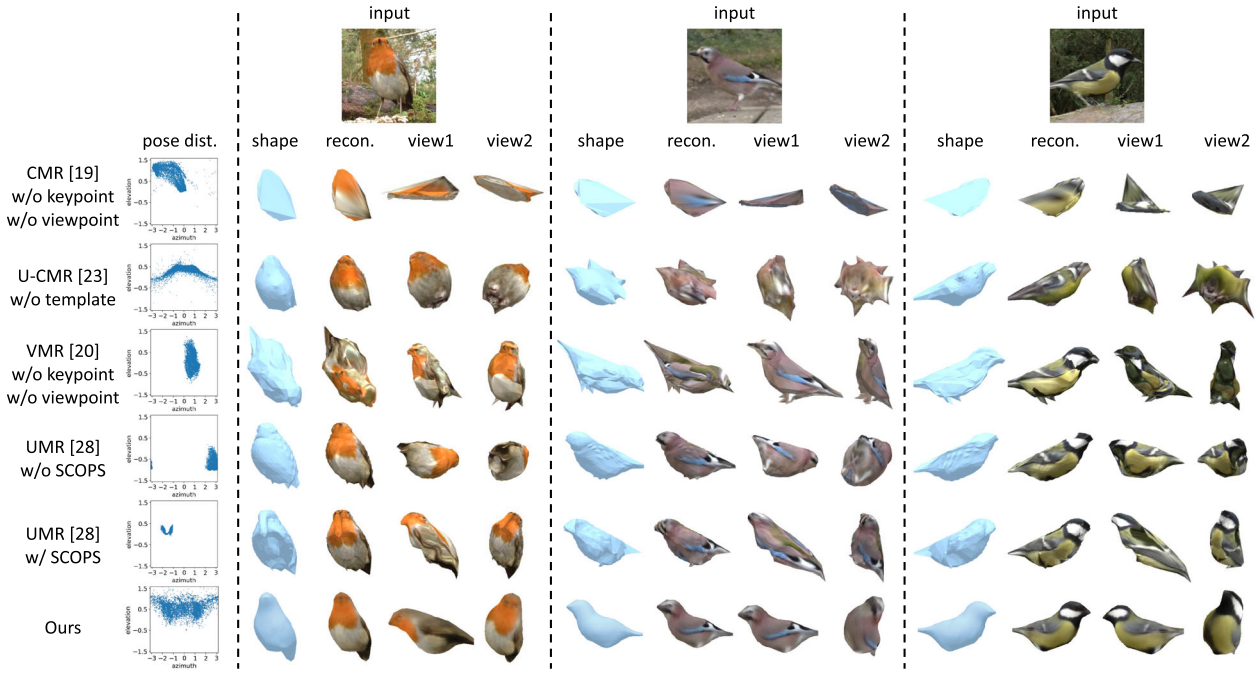

Fig. 5 Visual comparison. We compare to state-of-the-art methods trained without external geometric supervision in the form of 2D keypoints, viewpoint, or template shape. As UMR leverages weak-supervision using part segmentation maps from SCOPS (Hung et al., 2019), we show versions trained with and without SCOPS. Our method consistently reconstructs reasonable 3D shapes and the predictions cover full 360-degree (azimuth) view, whereas other methods produce poor reconstructions and their viewpoint predictions collapse to only a limited range with the exception of U-CMR. Other methods, except for U-CMR directly copy the texture from the input image using texture flow. Hence, although the texture appears sharper from the input view, they are often incorrect as seen from other views. See the sup. mat. for extended results

by PointRend (Kirillov et al., 2020) as the pseudo ground-truth. This evaluates the shape and viewpoint quality as the object from a past frame is projected to a future frame which can only align when both shape and pose are estimated correctly, but cannot account for non-rigid deformation between frames. For each test sequence, we predict the shape at frame t and render the object mask from the pose at frame $t + \Delta t$ with an offset Δt of 0, 5 and 20 frames. We then compute the mean Intersection over Union (mIoU) between the rendered masks and the ground-truth masks at $t + \Delta t$. Table 3 summarizes the results, which suggest that our model achieves both better shape reconstruction and viewpoint consistency. We also compute the metrics on our model with frame-specific deformations predicted at frame $t + \Delta t$ applied to the shape predicted at frame t. This further improves the mask reprojection IoU, confirming that our model learns correct frame-specific deformations. Other methods tend to overfit the shape to the image, resulting in a larger decrease in reprojection accuracy with increasing Δt.

We also compare the distribution of estimated viewpoints/object poses by plotting the elevation and azimuth predicted on the test set in Fig. 5. Our method is able to learn the full azimuth range, while other methods, with the exception of U-CMR, only predict limited range of views (azimuth) without additional geometric supervision.

Qualitative Comparisons

Figure 5 shows a qualitative comparison of different methods. When methods relying on more geometric supervision (CMR, U-CMR, VMR) are trained without this learning signal, they fail to produce reasonable shape reconstructions. UMR trained without SCOPS part segmentations overfits to the input views producing inaccurate 3D shapes. Our method reconstructs accurate shape and pose, despite not using keypoint or template supervision. We refer the reader to the sup. mat. for more results. Note that our model is trained on 128×128 images, whereas other methods train on 256×256 images and, except U-CMR, sample the texture directly from the input image, explaining the difference in the texture quality.

On Horse Video Dataset

For horses, we compare qualitatively with LASR (Yang et al., 2021a) in Fig. 6, which is an optimization-based method for single video sequences. While their reconstruction appears to be convincing in the original viewpoint, the actual mesh often does not resemble the shape of a horse. Running LASR on such a sequence takes over four hours.

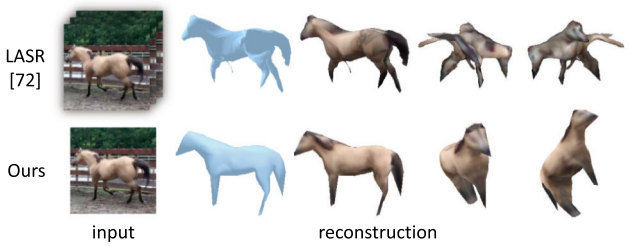

Fig. 6 Comparison with LASR (Yang et al., 2021a). While the rendering in the original viewpoint looks convincing, the shape produced by LASR is distorted and does not resemble the actual shape of a horse. Since our method trains on multiple sequences it can learn a consistent shape

4.4 Ablation and Analysis

We ablate the different components of our method quantitatively on our toy bird dataset in Table 4 and Fig. 7. We find that all components are necessary for the final performance. The pose distribution in Table 4 shows that the model only learns the full 360-degree (azimuth) view of the object when all components are active. Especially the two-view-ambiguity resolution and the shape symmetry are important to learn the pose while video training helps to discover the backside of the object. Without a good pose prediction the reconstructions look reasonable in the input view, reveal to be degenerate from other directions.

The model without symmetry produces unrealistic shapes indicating that symmetry is a useful prior, even when learning deformable shapes. Similarly, the shape prior is important to discover fine details (*e.g.* beak and tail) that are not visible in every image. The full model predicts a full range of viewpoints (Fig. 7) and the most consistent shape (Table 4).

We train another model without the learned category prior shape, predicting individual shapes for each bird. The resulting reconstructions are inconsistent across different instances, shown in Fig. 7. This suggests that the full model is able to leverage shape prior of the whole category, which is a major benefit of *learning* in a reconstruction pipeline.

Table 4 Ablation studies with 3D toy bird scans

	Chamfer Distance (cm)
Full model	**1.51** ±0.89
w/o front-back hypothesis	2.52 ±1.41
w/o symmetry	2.19 ±1.24
w/o video training	2.20 ±1.03
w/o learned prior shape	3.92 ±1.47

Every component of our model helps to improve the final performance

5 Limitations and Future Work

Our method still requires segmentation masks obtained from the off-the shelf model as supervision for training. Moreover, their quality affects the fidelity of our reconstructions. Thus, similar to comparable methods, our reconstructions do not capture fine details well, such as legs and the beak. The texture prediction sometimes results in low quality reconstructions especially when the input image is affected by motion blur. Currently, the predicted mesh is deformed from a sphere with fixed topology. Moreover, we have to handcraft a structure for various types of animals, for example different structures for horses (quadrupeds) and birds. How to automatically discover plausible bone structures is also an interesting question to explore for future work.

6 Conclusions

We have presented a method to learn articulated 3D representations of deformable objects from monocular videos without explicit geometric supervision, such as keypoints, viewpoint or template shapes. The resulting 3D meshes are temporally consistent and can be animated. The method can be trained from videos and only needs off-the-shelf object detection and

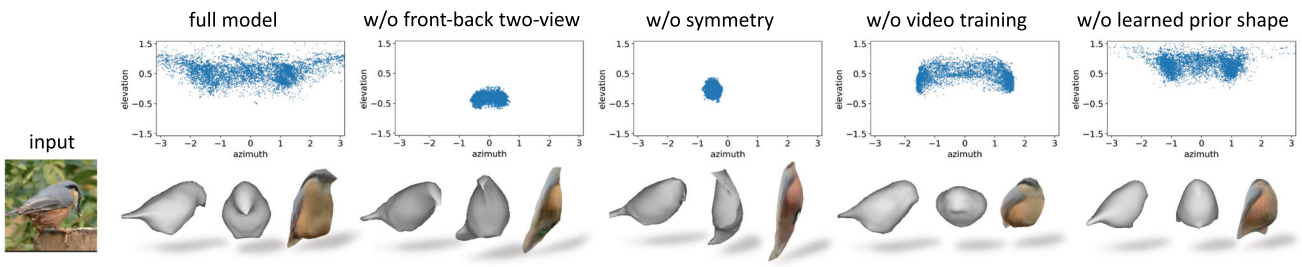

Fig. 7 Ablation studies. We train our model without some of the key components and plot the distribution of the predicted poses. Without 2-view ambiguity resolution or symmetry constraint, the pose prediction collapses. Video training and learning a shape prior also help improve the poses and shapes

optical flow models for preprocessing. For reproducibility, comparison and benchmarking, the dataset, code and models will be released with the paper.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-023-01819-5.

Acknowledgements We thank Zirui Wang for insightful discussions and Xueting Li for sharing the code for VMR with us. Shangzhe Wu is supported by Meta Research. Tomas Jakab is supported by Clarendon Scholarship. Christian Rupprecht is supported by Innovate UK (project 71653) on behalf of UK Research and Innovation (UKRI) and the Department of Engineering Science at the University of Oxford. Andrea Vedaldi is supported by EPSRC Grant (Visual AI EP/T028572/1).

References

Apple Reality Kit Object Capture API. https://developer.apple.com/augmented-reality/object-capture/. Accessed: 2021-10

Arnab, A., Doersch, C., & Zisserman, A. (2019). Exploiting temporal context for 3d human pose estimation in the wild. In: CVPR.

Besl, P. J., & McKay, N. D. (1992). A method for registration of 3-D shapes. *IEEE TPAMI 14*(2).

Chan, E., Monteiro, M., Kellnhofer, P., Wu, J., & Wetzstein, G. (2021). pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: CVPR.

Chen, C., Tyagi, A., Agrawal, A., Drover, D., MV, R., Stojanov, S., & Rehg, J. M. (2019). Unsupervised 3d pose estimation with geometric self-supervision. In: CVPR.

Choy, C.B., Xu, D., Gwak, J., Chen, K., & Savarese, S. (2016). 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: ECCV.

Doersch, C., & Zisserman, A. (2019). Sim2real transfer learning for 3d human pose estimation: motion to the rescue. NeurIPS.

Faugeras, O., & Luong, Q.-T. (2001). *The Geometry of Multiple Images*. London: MIT Press.

Feng, Y., Wu, F., Shao, X., Wang, Y., & Zhou, X. (2018). Joint 3d face reconstruction and dense alignment with position map regression network. In: ECCV.

Fernandez-Labrador, C., Chhatkuli, A., Paudel, D. P., Guerrero, J. J., Demonceaux, C., & Gool, L. V. (2020). Unsupervised learning of category-specific symmetric 3D keypoints from point sets. In: ECCV.

Gafni, G., Thies, J., Zollhöfer, M., & Nießner, M. (2021). Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: CVPR.

Garg, R., Kumar, B. G. V., Carneiro, G., & Reid, I. D. (2016). Unsupervised CNN for single view depth estimation: Geometry to the rescue. In: ECCV.

Girdhar, R., Fouhey, D., Rodriguez, M., & Gupta, A. (2016). Learning a predictable and generative vector representation for objects. In: ECCV.

Gkioxari, G., Malik, J., & Johnson, J. (2019). Mesh r-cnn. In: ICCV.

Goel, S., Kanazawa, A., & Malik, J. (2020). Shape and viewpoint without keypoints. In: ECCV.

Groueix, T., Fisher, M., Kim, V. G., Russell, B., & Aubry, M. (2018). AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In: CVPR.

Hartley, R., & Zisserman, A. (2004). *Multiple View Geometry in Computer Vision* (2nd ed.). Cambridge: Cambridge University Press. ISBN: 0521540518.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. B. (2017). Mask R-CNN. In: ICCV.

Henderson, P., & Ferrari, V. (2019). Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. IJCV, 1–20.

Henzler, P., Mitra, N. J., & Ritschel, T. (2019). Escaping plato's cave using adversarial training: 3D shape from unstructured 2D image collections. In: ICCV.

Henzler, P., Reizenstein, J., Labatut, P., Shapovalov, R., Ritschel, T., Vedaldi, A., & Novotny, D. (2021). Unsupervised learning of 3d object categories from videos in the wild. In: CVPR.

Hung, W., Jampani, V., Liu, S., Molchanov, P., Yang, M., & Kautz, J. (2019). SCOPS: self-supervised co-part segmentation. In: CVPR.

Kanazawa, A., Black, M. J., Jacobs, D. W., & Malik, J. (2018). End-to-end recovery of human shape and pose. In: CVPR.

Kanazawa, A., Tulsiani, S., Efros, A. A., & Malik, J. (2018). Learning category-specific mesh reconstruction from image collections. In: ECCV.

Kanazawa, A., Zhang, J. Y., Felsen, P., & Malik, J. (2019). Learning 3d human dynamics from video. In: CVPR.

Kato, H., & Harada, T. (2019). Learning view priors for single-view 3d reconstruction. In: CVPR.

Kato, H., Ushiku, Y., & Harada, T. (2018). Neural 3d mesh renderer. In: CVPR.

Kirillov, A., Wu, Y., He, K., & Girshick, R. (2020). PointRend: Image Segmentation as Rendering. In: CVPR.

Kokkinos, F., & Kokkinos, I. (2021). Learning monocular 3d reconstruction of articulated categories from motion. In: CVPR.

Kokkinos, F., & Kokkinos, I. (2021). To the point: Correspondence-driven monocular 3d category reconstruction. In: NeurIPS.

Kudo, Y., Ogaki, K., Matsui, Y., & Odagiri, Y. (2018). Unsupervised adversarial learning of 3D human pose from 2D joint locations. arXiv preprint arXiv:1803.08244.

Kulkarni, N., Gupta, A., & Tulsiani, S. (2019). Canonical surface mapping via geometric cycle consistency. In: ICCV.

Kulkarni, N., Gupta, A., Fouhey, D. F., & Tulsiani, S. (2020). Articulation-aware canonical surface mapping. In: CVPR, pp. 449–458.

Li, X., Liu, S., Kim, K., Mello, S. D., Jampani, V., Yang, M., & Kautz, J. (2020). Self-supervised single-view 3D reconstruction via semantic consistency. In: ECCV.

Li, X., Liu, S., Mello, S.D., Kim, K., Wang, X., Yang, M., & Kautz, J. (2020). Online adaptation for consistent mesh reconstruction in the wild. In: NeurIPS.

Liu, S., Li, T., Chen, W., & Li, H. (2019). Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In: ICCV.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: a skinned multi-person linear model. ACM TOG.

Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., & Geiger, A. (2019). Occupancy networks: Learning 3d reconstruction in function space. In: CVPR.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. In: ECCV.
Neverova, N., Novotný, D., & Vedaldi, A. (2020). Continuous surface embeddings. In: NeurIPS.
Newcombe, R. A., Fox, D., & Seitz, S. M. (2015). DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In: CVPR.
Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohli, P., Shotton, J., Hodges, S., & Fitzgibbon, A. W. (2011). KinectFusion: Real-time dense surface mapping and tracking. In: Proc. ISMAR.
Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., & Yang, Y.-L. (2019). Hologan: Unsupervised learning of 3d representations from natural images. In: ICCV.
Nguyen-Phuoc, T., Richardt, C., Mai, L., Yang, Y., & Mitra, N. J. (2020). Blockgan: Learning 3d object-aware scene representations from unlabelled images. In: NeurIPS.
Niemeyer, M., & Geiger, A. (2021). Giraffe: Representing scenes as compositional generative neural feature fields. In: CVPR.
Niemeyer, M., Mescheder, L., Oechsle, M., & Geiger, A. (2020). Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: CVPR.
Noguchi, A., Sun, X., Lin, S., & Harada, T. (2021). Neural articulated radiance field. In: ICCV.
Novotný, D., Larlus, D., & Vedaldi, A. (2017). Learning 3D object categories by looking around them. In: ICCV.
Oechsle, M., Peng, S., & Geiger, A. (2021). Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: ICCV.
Pan, X., Dai, B., Liu, Z., Loy, C. C., & Luo, P. (2021). Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. In: ICLR.
Park, J.J., Florence, P., Straub, J., Newcombe, R., & Lovegrove, S. (2019). Deepsdf: Learning continuous signed distance functions for shape representation. In: CVPR.
Park, K., Sinha, U., Barron, J. T., Bouaziz, S., Goldman, D. B., Seitz, S. M., & Martin-Brualla, R. (2021). Nerfies: Deformable neural radiance fields. In: ICCV.
Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. In: Advanced Video and Signal Based Surveillance.
Pumarola, A., Corona, E., Pons-Moll, G., & Moreno-Noguer, F. (2020). D-NeRF: Neural Radiance Fields for Dynamic Scenes. In: CVPR.
Raj, A., Zollhoefer, M., Simon, T., Saragih, J., Saito, S., Hays, J., & Lombardi, S. (2021). Pva: Pixel-aligned volumetric avatars. In: CVPR.
Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.-Y., Johnson, J., & Gkioxari, G. (2020). Accelerating 3d deep learning with pytorch3d. arXiv preprint arXiv:2007.08501
Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., & Novotny, D. (2021). Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: ICCV, pp. 10901–10911.
Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., & Li, H. (2019). Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: ICCV.
Sanyal, S., Bolkart, T., Feng, H., & Black, M. (2019). Learning to regress 3D face shape and expression from an image without 3D supervision. In: CVPR.
Schwarz, K., Liao, Y., Niemeyer, M., & Geiger, A. (2020). Graf: Generative radiance fields for 3d-aware image synthesis. In: NeurIPS.
Sorkine, O., & Alexa, M. (2007). As-rigid-as-possible surface modeling. In: Symposium on Geometry Processing, vol. 4, pp. 109–116.
Teed, Z., & Deng, J. (2020). RAFT: recurrent all-pairs field transforms for optical flow. In: ECCV.
Thewlis, J., Bilen, H., & Vedaldi, A. (2018). Modelling and unsupervised learning of symmetric deformable object categories. In: NeurIPS.
Tran, L., & Liu, X. (2019). On learning 3d face morphable model from in-the-wild images. *IEEE TPAMI, 43*(1), 157–171.
Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., & Theobalt, C. (2021). Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: ICCV.
Tulsiani, S., Kulkarni, N., & Gupta, A. (2020). Implicit mesh reconstruction from unannotated image collections. arXiv preprint arXiv:2007.08504.
Tung, H., Tung, H., Yumer, E., & Fragkiadaki, K. (2017). Self-supervised learning of motion capture. In: NIPS.
Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.
Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., & Wang, W. (2021). Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In: NeurIPS.
Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., & Jiang, Y.-G. (2018). Pixel2mesh: Generating 3d mesh models from single rgb images. In: ECCV.
Wu, S., Makadia, A., Wu, J., Snavely, N., Tucker, R., & Kanazawa, A. (2021). De-rendering the world's revolutionary artefacts. In: CVPR.
Wu, S., Rupprecht, C., & Vedaldi, A. (2020). Unsupervised learning of probably symmetric deformable 3D objects from images in the wild. In: CVPR. https://elliottwu.com/projects/unsup3d/.
Wu, J., Zhang, C., Xue, T., Freeman, W. T., & Tenenbaum, J. B. (2016). Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: NeurIPS, pp. 82–90.
Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Chang, H., Ramanan, D., Freeman, W.T., & Liu, C. (2021). LASR: Learning articulated shape reconstruction from a monocular video. In: CVPR.
Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Liu, C., & Ramanan, D. (2021). Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. In: NeurIPS.
Yang, G., Vo, M., Natalia, N., Ramanan, D., Andrea, V., & Hanbyul, J. (2022). Banmo: Building animatable 3d neural models from many casual videos. In: CVPR.
Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Ronen, B., & Lipman, Y. (2020). Multiview neural surface reconstruction by disentangling geometry and appearance. In: NeurIPS.
Ye, Y., Tulsiani, S., & Gupta, A. (2021). Shelf-supervised mesh prediction in the wild. In: CVPR.
Zhang, Y., Chen, W., Ling, H., Gao, J., Zhang, Y., Torralba, A., & Fidler, S. (2021). Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. In: ICLR.
Zhang, J. Y., Felsen, P., Kanazawa, A., & Malik, J. (2019). Predicting 3d human dynamics from video. In: ICCV.
Zhou, T., Brown, M., Snavely, N., & Lowe, D. G. (2017). Unsupervised learning of depth and ego-motion from video. In: CVPR.
Zuffi, S., Kanazawa, A., Berger-Wolf, T., & Black, M. J. (2019). Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In: ICCV.
Zuffi, S., Kanazawa, A., Jacobs, D., & Black, M. J. (2017). 3D menagerie: Modeling the 3D shape and pose of animals. In: CVPR.