Bootstrapping Articulated 3D Reconstruction from 2D Image Collections

Jakub Zadrozny, Oisin Mac Aodha, and Hakan Bilen

University of Edinburgh
https://jakubzadrozny.github.io/bat3r

Abstract. 3D reconstruction of articulated objects from a single image is challenging because large training datasets with paired image and 3D supervision are difficult to obtain. Recent point map–based methods achieve strong performance but rely on synthetic datasets rendered from manually created articulated 3D assets with carefully curated pose distributions. While camera viewpoints can be easily sampled, generating realistic object articulations remains costly and labor-intensive. We propose a training framework that reduces this requirement by leveraging unannotated 2D images collections with only a single rigged canonical mesh per category. Starting from a weak 3D shape predictor trained on canonical-pose renders, we iteratively estimate object articulation and camera pose by fitting the mesh to predicted point maps. The recovered articulations and viewpoints are then used to render updated synthetic training data, progressively improving the predictor. Despite using substantially weaker 3D supervision, our models achieve performance comparable with DualPM, which requires manually curated articulated training datasets.

Keywords: 3D Reconstruction · Weak Supervision · Articulation

1 Introduction

Estimating the 3D shape of an object from a single image is an inherently ill-posed task in computer vision, as multiple 3D structures can correspond to the same 2D observation. The challenge becomes even more pronounced in the presence of occlusions, whether self-occlusions or those caused by other objects in the scene, which leave only partial information about the object's geometry. Addressing this problem therefore requires reasoning about the complete 3D structure from limited visual evidence. This ability to infer and understand such complex 3D shapes is fundamental for next-generation embodied systems that must perceive, reason about, and interact with objects in the physical world.

Previous works addressing this problem typically rely on incorporating priors to constrain the otherwise ambiguous 3D reconstruction space, ensuring that predicted shapes are both plausible and consistent with the input image. Such priors have been introduced in several ways: through morphable 3D shape templates [23, 48], learned from 2D image collections [1, 19, 37] or by encoding shape

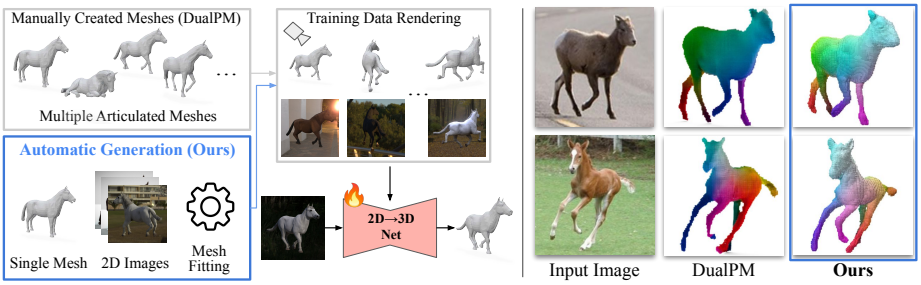

Fig. 1: While effective, supervised approaches for 3D shape estimation from 2D images rely on the availability of large manually crafted and curated 3D shape datasets depicting objects of interest in a wide variety of poses (top left). We introduce a new approach that only requires a single mesh per category and an image collection as input (bottom left). From this, we are able to automatically generate paired training data, *i.e.,* 2D images and corresponding 3D point clouds, that can be used to train supervised reconstruction methods. Our 3D prediction results are comparable to methods that are trained on richer manually created 3D articulated data, *e.g.,* DualPM [12] (right).

knowledge into neural networks with supervised 3D data [12, 39]. Among these approaches, supervised models trained on paired input images and corresponding 3D data (*e.g.,* meshes) have demonstrated particularly strong performance [4].

However, like most supervised approaches, the success of these methods is heavily reliant on the quality and diversity of available training data. While recent efforts have expanded the availability of synthetic 3D assets through large-scale dataset curation [5, 6, 18], these collections still suffer from limitations in realism and diversity. Moreover, the manual curation required to build larger, high-quality datasets is prohibitively expensive. This issue is especially pronounced for object-centric 3D reconstruction of deformable and articulated objects (*e.g.,* animals), which require substantial amounts of diverse 3D data for effective supervised training. Outside of human-centric domains [2, 3, 33, 41], there remains a scarcity of 3D datasets for other complex articulated objects.

While obtaining large-scale articulated 3D data for a specific object category is challenging, as discussed above, single meshes depicting the category in a rest (*i.e.,* canonical or neutral) pose are often readily available. Motivated by this observation, we ask whether it is possible to train a model to estimate the 3D shape of an object from a single image given only a single 3D model of that category and a set of 2D images at training time? In doing so, we would significantly decrease the amount of manual curation required to generate such datasets and provide a scalable pathway to training models for categories that have limited 3D data available.

At first glance, this appears infeasible as a single mesh cannot capture the wide range of pose and shape variation observed in real-world images. Nevertheless, we observe that training from rendered images obtained from a single 3D mesh can provide a strong initialization for basic shape and camera viewpoint estimation (see Fig. 1 for an overview). Building on this insight, we propose an

iterative data generation procedure that progressively augments the training set with predicted paired 2D and 3D training instances that exhibit progressively more complex poses. In doing so, we demonstrate that it is possible to train the recent state-of-the-art DualPM [12] 3D shape estimation method on significantly weaker 3D data yet still obtain comparable performance on challenging quadruped animal categories.

We make the following contributions: (i) We propose a new approach, Bootstrapping ArTiculated 3D Reconstruction (BAT3R), for automatically generating paired 2D and 3D training data that can be used to train category-centric 2D to 3D shape prediction models while foregoing the need for expensive manually created 3D articulated training data. (ii) Despite being trained on significantly weaker 3D data, we demonstrate that our approach results in strong reconstruction performance. Furthermore, we show that the iterative application of our method improves performance, closing the gap between our approach and fully supervised baselines.

2 Related work

There is a large body of work in computer vision that attempts to reconstruct a 3D representation of a scene given a set of images depicting the scene, *e.g.,* [13,24, 26,31,35]. In this work, our focus is on a different type of 3D reconstruction task, that of reconstructing deformable objects from unstructured image collections.

Unsupervised reconstruction. There is a family of unsupervised methods that have been proposed that can estimate the 3D shape of a *rigid* object category given a collection of images [8, 11, 17, 25, 34]. Unlike conventional structure from motion methods that assume the provided images are from different viewpoints of same object instance or scene at similar points in time, these methods instead take an unstructured set of images as input, *i.e.,* a set of images depicting the same object category, but not necessarily the same exact instance. The methods are unsupervised in the sense that no ground-truth 3D supervision is available, but they instead make use of auxiliary 2D supervision such as foreground segmentation masks, semantic part segmentation masks, object keypoints, camera poses, *etc*. Successive methods have managed to remove some of the supervision signal required either partially (*e.g.,* by removing the need for keypoint annotations [8]) or completely [25]. Due the rigid object assumptions made by these methods, they cannot estimate plausible 3D shapes for articulated/deformable objects.

Techniques have been proposed to address the more challenging non-rigid case whereby objects may deform or articulate. Given a collection of images and a skeleton, LASSIE [42] introduced a test-time optimization approach to reconstruct articulated categories. Later this was extended such that the skeleton could also be estimated from the images [43]. An alternative approach is to train a neural network to directly predict the articulated 3D shape. Skeleton-based priors have also been demonstrated to be effective in this setting as in

MagicPony [37] or 3D-Fauna [19], but other priors such as motion cues obtained from video sequences, as used in DOVE [36], are also effective. SAOR [1] does not require any 3D or video supervision, instead relying on the assumption that objects can be decomposed into a set of articulated parts which are learned during training. While impressive, the results of these unsupervised methods are still inferior to techniques that are trained with full 3D supervision.

Supervised reconstruction. The above unsupervised methods broadly do not make use of any detailed 3D shape information during training. However, when available, this can impose a strong regularization signal ensuring that model predictions stay on the manifold of plausible 3D shapes. One approach is to align multiple 3D meshes of different instances of the same object category into a common model which captures the shape variation within the category. This has been shown to be highly effective for humans [23], quadrupeds [48] more generally, and specific species such as dogs [29,30] and horses [49]. However, the 3D alignment step is non-trivial and limits this approach to settings where a large number of 3D meshes containing instances of the same topology are available.

When large quantities of paired image and 3D data are available, it is possible to directly apply supervised learning techniques to learn the mapping from images to 3D, *e.g.,* [47]. Additionally, priors learned from deep generative (*e.g.,* diffusion-based) models have also been shown to be effective at assisting with 3D extraction [21, 22, 39]. These supervised approaches require training data, but advances in rendering and the availability of large 3D asset collections have enabled the creation of paired datasets in the case of a limited number of deformable categories such as animals, *e.g.,* [9, 27, 40]. However, the generation of such datasets relies on the availability of 3D assets in diverse poses.

Inspired by point-based representations developed in the context of 3D reconstruction networks [35] and canonical surface mapping [15, 16], DualPM [12] introduced a supervised approach for viewpoint invariant point map prediction for deformable object categories. It leverages large synthetically rendered images and 3D point clouds to train a supervised model that can predict both the 3D position of visible and occluded parts of the object via a multi-layered output representation. DualPM produces state-of-the-art reconstruction results for quadrupeds. However, like the supervised methods previously discussed, it requires 3D training data depicting objects in diverse poses. In this work, we propose a new approach for automatic dataset generation that overcomes this requirement for diverse data. Specifically, we show that it is possible to generate rich training data via an iterative procedure that only requires a single rigged canonically posed mesh as input.

3 Method

Given a collection of RGB images $\{I_n\}_{n=1}^{N}$ depicting different instances of an articulated object category (*e.g.,* horses), our goal is to train a single-image predictor Φ that reconstructs the object's 3D geometry. At training time, we

assume we have access to a single canonical mesh M^c representing the object in its rest pose. We also assume that M^c is animation-ready, *i.e.,* rigged and skinned with an articulation skeleton. Such meshes can be created manually or obtained using recent automatic rigging methods [32,45]. Importantly, we do not require additional instances of this mesh in different articulated poses.

For the image collection, we assume foreground segmentation masks are available, which can be obtained using off-the-shelf segmentation methods [14, 28]. The images require neither camera pose annotations nor 3D supervision. We also do not assume prior knowledge of the distribution of valid object articulations, and only require a coarse prior over camera viewpoints. Both object articulations and camera poses are estimated automatically during training. We use the DualPM point-map representation [12] for the single-image predictor, which we summarize next.

3.1 Preliminaries: Dual Point Maps (DualPM)

Our main contribution is a framework for automatically generating paired 2D–3D training data for articulated image-to-3D reconstruction. For the reconstruction model, we build on DualPM [12], which predicts two coupled, layered 3D point maps from a single image.

Given an input image I, for each foreground pixel u, the network predicts K intersection points between a ray passing through pixel u and the object surface (capturing both visible and occluded parts). Each intersection point at layer k is represented by a posed point $P_k(u) \in \mathbb{R}^3$ and a canonical point $Q_k(u) \in \mathbb{R}^3$. The posed point is expressed in the camera coordinate frame, while the canonical point is expressed in the object's rest-pose coordinate system. Together, $(P_k(u), Q_k(u))$ describe the same physical surface location in two coordinate frames. The posed map therefore provides a 3D reconstruction of the object in the camera coordinate frame, while the canonical map establishes dense correspondences mapping canonical points to posed points. This explicit canonical-to-posed mapping is the core mechanism driving our mesh fitting stage.

DualPM [12] is a fully supervised approach and it is trained using synthetically rendered images depicting articulated meshes where ground-truth posed and canonical coordinates are available for each surface point (both visible and occluded). We let $\hat{P}_k(u)$ and $\hat{Q}_k(u)$ denote the predicted posed and canonical points, respectively, and F denote the set of masked object (foreground) pixels. The posed point map is trained using the confidence-weighted regression loss:

$$\mathcal{L}_P = \frac{1}{K|F|} \sum_{u \in F} \sum_{1 \le k \le K} c_{P_k}(u) \, \|\hat{P}_k(u) - P_k(u)\|_2^2 - \alpha \log c_{P_k}(u), \tag{1}$$

where $P_k(u)$ and $Q_k(u)$ denote the ground truth posed and canonical points, $c_{P_k}(u)$ denotes a predicted confidence for the posed point estimate. The total training objective is $\mathcal{L}_{DualPM} = \mathcal{L}_P + \mathcal{L}_Q$, where an analogous loss to $\mathcal{L}_P$ is used for $\mathcal{L}_Q$.

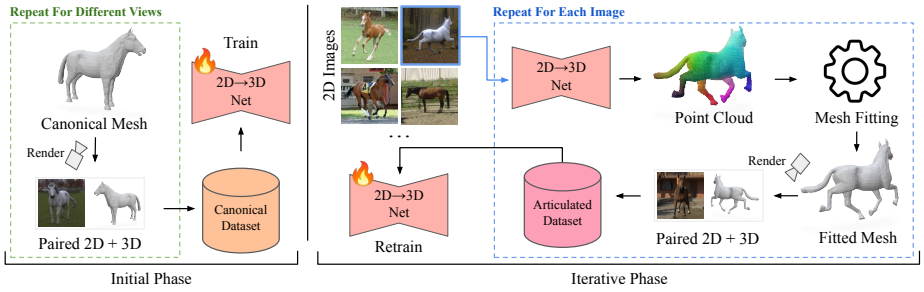

Fig. 2: Overview of our bootstrapping approach BAT3R. Left: We render the canonical rigged mesh from sampled camera viewpoints to create an initial paired 2D–3D training set, which is used to train the initial 2D to 3D predictor Φ_0. Right: Given an image collection, the current predictor Φ_i predicts posed and canonical point maps. We fit the canonical mesh to these predictions, recover an articulated mesh and camera viewpoint, and render the result to obtain new synthetic training pairs with exact point-map supervision. The predictor is then retrained on this generated dataset, yielding Φ_{i+1}. The process is repeated for multiple iterations.

3.2 BAT3R: Bootstrapping ArTiculated 3D Reconstruction

The original DualPM framework requires a training collection of 2D images depicting the object of interest in various articulated poses, paired with corresponding ground-truth layered 3D point maps with canonical correspondences. In practice, this paired data is generated synthetically via layered rasterization of articulated 3D meshes. However, acquiring such high-quality datasets is highly non-trivial as creating realistic articulation data for rigged meshes is a challenging, time-consuming, and expensive process that typically demands the expertise of professional 3D artists. Furthermore, achieving comprehensive coverage of natural articulations is critical, as the predictive performance and generalization capabilities of the learned model are inherently limited by the pose and articulation diversity present within the training data. We propose an alternative solution that requires significantly less 3D supervision. Our method alternates between two steps: fitting the canonical mesh to predictions on the 2D image collection, and using the fitted meshes to render new paired 2D–3D training data.

We assume we have access to an initial, potentially weak, 3D predictor Φ_i, which can be initialized by training the predictor on image and 3D data rendered from a single canonical mesh (*i.e.,* depicting a single pose). The high-level idea behind our approach is that we can use the initial predictor to generate plausible predictions for the supplied image collection and then make use of mesh alignment methods to fit the canonical mesh to the predicted point clouds to obtain realistic object poses that can be used to generate improved ground-truth 3D training data. Below we detail a single iteration of our proposed refinement

procedure, designed to yield an improved predictor, Φ_{i+1}. An illustration of our method, BAT3R, is depicted in Fig. 2.

We first employ the current model Φ_i to infer the dual point clouds (derived from the dual point maps), *i.e.,* the posed point cloud P and the corresponding canonical point cloud Q, for each image I in the training collection. Subsequently, we optimize for the best articulation of the provided canonical 3D mesh to the predicted posed point cloud for each instance. This fitting process is facilitated by the semantic correspondences provided by the dual point cloud representation (see Sec. 3.1). Although the geometric fidelity of the predicted point clouds may be imperfect (particularly during early stages of the refinement), their primary function is merely to guide the mesh articulation toward the target pose depicted in the input image. A key observation motivating our approach is that exact ground-truth 3D annotations for the input images are not required to iteratively improve the predictor. Instead, we rely on articulating the canonical mesh to fit the imperfect predictions, which inherently provides a fully accurate and self-consistent 3D structure for supervision. Furthermore, as a natural by-product of our mesh alignment process, we also recover the camera poses corresponding to the input images. Finally, these articulated meshes and estimated camera poses are utilized to render a novel, self-annotated training set, which serves as the supervisory signal to train the subsequent model iteration, Φ_{i+1}.

Initialization. In the most general case, we begin the process assuming no knowledge about the distribution of plausible 3D articulations and simply train a naïve initial model to serve as Φ_0, *i.e.,* Φ_i for the first iteration where $i = 0$. To this end, we automatically create a dataset of 2D renders of the canonical mesh M^c (*i.e.,* a 3D mesh of the object in a rest pose) from many randomly sampled camera viewpoints. The distribution used to sample camera viewpoints should in general resemble the distribution that is expected to be encountered at test time. To complete the rendering pipeline we require an object texture, which can be hand-crafted or AI-generated, and a environment map which we sample from a collection of publicly available assets. This dataset of images of the object in the canonical pose from various viewpoints, with known 3D per-image shape, serves as our initial paired training data.

Inference and mesh fitting. Our mesh fitting algorithm employs a gradient-based optimization loop to articulate and align the canonical 3D mesh to the predicted dual point clouds. The dual point map representation lends itself particularly well to this purpose by establishing explicit dense correspondences between the posed and the canonical states. Specifically, let $V^c = \{v_j^c\}$ denote the vertices of the canonical mesh M^c, and let $P = \{p_i\}$ and $Q = \{q_i\}$ represent the predicted (using Φ_i) posed and canonical point clouds for a specific image I, respectively. First, each point indexed by i in the dual point map is assigned to the mesh vertex that minimizes the distances in the canonical space, *i.e.,* $j^*(i) = \arg\min_j \|q_i - v_j^c\|_2$. Intuitively, during the fitting process, the cor-

responding predicted posed point p_i acts as a target attractor pulling on the articulated mesh vertex.

Furthermore, when fitting the mesh to point clouds predicted from in-the-wild images, the template mesh M^c will inherently exhibit anatomical discrepancies compared to the observed instance. To prevent the optimization from overfitting to these instance-specific structural variations, we subsample the predicted dual point clouds to a sparse set of 200 control points. To ensure these control points are sampled uniformly along the surface of the object, we employ a topology-aware farthest point sampling strategy. We define a surface-grounded distance metric $D(q_n, q_m)$ between any two predicted canonical points $q_n, q_m \in Q$. Let $j^*(n)$ denote the index of the mesh vertex assigned to q_n, and let $d_{geo}(\cdot, \cdot)$ represent the geodesic distance along the canonical mesh surface. We formulate the distance as the path length bridging the two points via the mesh surface:

$$D(q_n, q_m) = \|q_n - v^c_{j^*(n)}\|_2 + d_{geo}\left(v^c_{j^*(n)}, v^c_{j^*(m)}\right) + \|v^c_{j^*(m)} - q_m\|_2 \quad (2)$$

For ease of notation, in all subsequent sections, we assume P and Q refer directly to these subsampled control point clouds.

The optimization process begins by determining an optimal global rigid transformation T^g (comprising scaling, rotation, and translation) using the Kabsch algorithm [10], to align the canonical mesh (defined in the rest-pose coordinate frame) with the predicted posed point cloud P (residing in the camera space). Following this global alignment, we optimize the articulation parameters θ, defined as the excess joint rotations relative to the rest pose, alongside per-bone (scalar) scale parameters S to account for instance-specific anatomical proportions (*e.g.,*, variations in limb length). The state of the posed mesh is thus a function of the articulation, bone scales, and global rigid transform, yielding the posed vertices $V^P(\theta, S, T^g) = \{v^p_j(\theta, S, T^g)\}$.

To guide this process, we define a loss function consisting of a data term and geometric regularization constraints. The data term, $\mathcal{L}_{data}$, measures the discrepancy between each predicted posed point p_i and its assigned articulated mesh vertex $v^p_{j^*(i)}$:

$$\mathcal{L}_{data} = \sum_{p_i \in P} \mathcal{H}_\delta\left(\|p_i - v^p_{j^*(i)}(\theta, S, T^g)\|_2\right). \quad (3)$$

To mitigate the influence of outliers and to accommodate natural deviations between the predicted point cloud and the specific instance's true shape, we employ a robust Huber loss $\mathcal{H}_\delta$ with $\delta = 0.1$ for this distance.

Our objective is not only to overfit the point cloud, but rather to find a close, reasonable alignment while preserving the intrinsic structural properties of the fitted mesh, which ensures the 3D predictor Φ learns a geometrically correct prior. To this end, we incorporate a set of regularization losses to address overfitting. First, inspired by physical energy minimization principles, we penalize the mean squared magnitude of the excess joint rotations ($\|\theta\|_2^2$), which encourages

smooth and gradual curvatures. Second, we introduce penalties against deviations in (local) volume $\mathcal{L}_{vol}$ and edge lengths $\mathcal{L}_{edge}$, between the canonical and articulated states, thereby preventing unrealistic stretching, shearing, or compression. Third, to ensure anatomical consistency, we regularize the bone scales via $\mathcal{L}_{scale}$, which enforces left-right symmetry and penalizes extreme deviations from the canonical proportions. Finally, when presented with challenging articulations, the 3D predictor tends to output noisy point clouds, especially early in the refinement process, which in turn results in self-intersections of the fitted meshes. To mitigate this effect, we apply a self-repulsion loss term, $\mathcal{L}_{rep}$, which explicitly penalizes configurations where vertices separated by a large geodesic distance on the mesh surface collapse to a small Euclidean distance in the articulated state, effectively preventing self-intersections. Crucially, to prevent these structural penalties from opposing the valid anatomical size variations captured by S, the geometric losses ($\mathcal{L}_{vol}, \mathcal{L}_{edge}, \mathcal{L}_{rep}$) are computed on the *unscaled* articulated mesh. We provide more details in the supplementary material.

The complete objective function is formulated as:

$$\mathcal{L} = \mathcal{L}_{data} + \lambda_{angle}\|\theta\|_2^2 + \lambda_{vol}\mathcal{L}_{vol} + \lambda_{edge}\mathcal{L}_{edge} + \lambda_{scale}\mathcal{L}_{scale} + \lambda_{rep}\mathcal{L}_{rep}. \quad (4)$$

The total loss $\mathcal{L}$ is jointly minimized using gradient descent with respect to the excess joint rotations θ, bone scales S, and the global rigid transform T^g, as articulating the mesh necessitates corrections to the global alignment. Importantly, the bone scales S act solely as auxiliary variables to absorb instance-specific anatomical variations for a more accurate data fit. Upon convergence, S is discarded, and we extract the unscaled articulated mesh to serve as the geometrically consistent 3D training target for the next iteration. Note that our regularization terms, apart from the angular penalty, which we assign a low weight λ_{angle}, do not hinder intricate yet natural deformations, as physically plausible articulations generally do not necessitate excessive stretching, compression, volume fluctuations, or self-intersections.

We empirically observe that our approach proves more effective at extracting accurate poses from lateral views compared to frontal and rear perspectives. To address this imbalance and boost performance on the challenging views, we employ a *viewpoint augmentation* strategy. Specifically, to form the training set for the next iteration, we render the fitted meshes from multiple different novel viewpoints (see supplement for details).

4 Experiments

4.1 Experimental setup

Datasets. For our synthetic data evaluation, we strictly follow the training splits provided by DualPM [12] for the horse, cow, and sheep categories. Crucially, we do not access the ground-truth articulations and camera poses. Instead, we recover the articulations and camera poses using our refinement procedure. For our real-world evaluation, we collect purely 2D in-the-wild images. For horses, we

utilize $\sim$11,000 real images from the MagicPony [37] dataset. For chimpanzees, we assemble a training collection of $\sim$19,000 images from AP-10K [44] and OpenApePose [7] (restricted to chimpanzees). Finally, for elephants, we assemble a training collection of $\sim$21,000 images comprising the *Wild Elephant Dataset* from Kaggle and frames extracted from YouTube videos using the Animal-in-Motion [46] pipeline.

3D assets and rendering. Our approach requires a canonical 3D mesh. For the horse, cow, and sheep categories, we utilize the exact canonical assets provided by DualPM. For chimps and elephants, we acquire publicly available meshes from the internet. While we utilize manually rigged meshes in this work, our pipeline is agnostic to the rigging source and can be paired with modern auto-rigging solutions [20, 45].

Evaluation protocol. Our evaluation spans both strict baseline comparisons and wider generalization tests. To this end, we utilize three quantitative test splits for the horse category: (1) *Animodel-Points* [9, 12]: the exact test split defined by DualPM; (2) *Horse-Robust*: a comprehensive, large-scale dataset of 15,000 samples (compared to $\sim$500 in Animodel-Points), where only the male horse template is available at training time, but the female template is used for evaluation; (3) *Horse-Mixed*: a standard in-distribution split (18,000 samples) where models are both trained and evaluated on a mix of male and female templates. We utilize this split for controlled ablation studies.

Evaluation metrics. Following the Animodel-Points benchmark protocol [9, 12], we evaluate the geometric accuracy of our predicted point clouds using the Root-Mean-Square (RMS) bi-directional Chamfer Distance (CD), reported in centimeters. We evaluate under both full rigid alignment (using Iterative Closest Point (ICP)) and model-view alignment (restricted to scale and translation), thereby penalizing inaccurate camera pose predictions. In the ablation studies, we report with more granularity, *i.e.,* the mean CD (mCD), the outlier ratio denoted as % (CD $>$ 5cm) which measures the fraction of test samples exceeding a 5cm error threshold, and the angular camera rotation error ΔR (°).

Baselines. We compare against category-specific models [9,15,19,37,39], including the state-of-the-art DualPM [12], which is trained with full 3D supervision. For real-world categories where 3D supervision is unavailable, we compare qualitatively against recent zero-shot 3D foundation models, namely SAM-3D [4] and Trellis.2 [38], which are evaluated in a zero-shot manner.

Implementation summary. We employ the DualPM architecture as our point map predictor. When training on synthetic images, we disable bone scaling, as these datasets lack the individual anatomical variation present in real-world samples. Furthermore, we observed that the original DualPM training regime becomes unstable when supervised by the mesh articulations fitted to noisy initial predictions, which inherently contain extreme outliers despite geometric regularization (particularly in the case of real-world images). To mitigate this instability, we adopt a modified optimization regime tailored for noisy, self-supervised targets (see the supplementary material for details). We typically run our refinement loop for four iterations (see Sec. 4.3 for analysis). Further imple-

Table 1: Quantitative evaluation on the Animodel-Points dataset. We report the Root Mean Square Error (RMS) of the bi-directional Chamfer Distance in centimeters (↓). Our methods in this comparison ('Canonical' and 'Ours (synt)') are trained on the synthetic images from DualPM following original splits. We report both the results with and without rotation alignment. Best results in **bold**, 2nd best underlined.

Method	RMS Chamfer Distance (cm) ↓			Model-view RMS Chamfer Dist. (cm) ↓		
	Horse	Cow	Sheep	Horse	Cow	Sheep
A-CSM [15]	11.75 ± 3.83	9.52 ± 2.41	9.24 ± 2.40	38.13 ± 13.89	33.51 ± 11.52	29.04 ± 9.35
MagicPony [37]	11.19 ± 3.08	10.29 ± 2.08	–	20.82 ± 13.04	25.39 ± 13.43	–
Farm3D [9]	11.34 ± 3.22	9.63 ± 2.02	11.01 ± 1.87	29.52 ± 15.73	21.34 ± 12.85	21.52 ± 9.84
3D-Fauna [19]	11.86 ± 3.03	10.54 ± 2.26	9.61 ± 2.15	15.70 ± 6.82	14.08 ± 4.20	12.24 ± 3.17
Trellis [39]	6.93 ± 4.13	6.80 ± 3.24	5.91 ± 2.93	36.82 ± 16.02	26.54 ± 13.14	26.56 ± 12.06
DualPM [12]	$\mathbf{4.30 \pm 1.50}$	$\mathbf{3.18 \pm 1.06}$	$\mathbf{3.30 \pm 1.20}$	$\mathbf{5.49 \pm 1.75}$	$\mathbf{4.03 \pm 2.11}$	$\mathbf{4.22 \pm 2.18}$
Canonical	7.88 ± 3.01	6.66 ± 2.31	6.62 ± 2.33	9.87 ± 4.55	7.85 ± 2.87	7.70 ± 2.66
Ours (synt)	$\underline{5.65 \pm 1.57}$	$\underline{4.17 \pm 1.15}$	$\underline{4.35 \pm 1.26}$	$\underline{7.73 \pm 2.06}$	$\underline{5.30 \pm 2.42}$	$\underline{5.04 \pm 1.71}$

Table 2: Quantitative evaluation on the Horse-Robust test split. We report the Root Mean Square Error of the bi-directional Chamfer Distance (RMS CD) and Model-view (MV) RMS CD. Both variants of our approach outperform large-scale foundation models despite using orders of magnitude less 3D supervision. Trellis.2 relies purely on best-effort post-hoc rotation alignment which occasionally fails, inflating its error. Best results in **bold**, 2nd best underlined.

Metric (cm) ↓	Trellis.2	SAM-3D	DualPM	Canonical	**Ours (synt)**	**Ours (real)**
RMS CD	8.92	6.42	**3.84**	7.84	$\underline{5.10}$	6.01
MV RMS CD	40.03	8.71	**5.42**	10.83	$\underline{7.92}$	8.37

mentation details, including texture generation, camera pose distributions, and mesh fitting parameters, are provided in the supplementary material.

4.2 Results

Training on synthetic images. In Tab. 1 we quantitatively compare our approach to existing single image 3D object reconstruction methods. We compare to both unsupervised and supervised methods, across the three object categories (horse, cow, and sheep) evaluated in DualPM. DualPM reports state-of-the-art performance on this task, and serves as a fully supervised upper bound for our approach. We observe that the model trained only on images rendered from the canonical mesh ('Canonical') is surprisingly effective, specifically on the comparatively easier sheep category which exhibits less articulation than horses. In general, our method, trained on the exact synthetic image splits provided by DualPM, approaches its performance while also outperforming the other baselines tested. See the supplementary material for a qualitative comparison with 3D-Fauna [19] and DualPM [12] on the *cow* and *sheep* categories.

Training on real images. Here we demonstrate the capabilities of our approach trained exclusively on real-world in-the-wild image collections. In Tab. 2 we compare the performance of our model trained on real-world images ('Ours (real)') to DualPM, our model trained only on images rendered from the canonical mesh ('Canonical'), our model trained on synthetic images ('Ours (synt)'), and large image-to-3D supervised foundation models [4, 38] on the wide *Horse-Robust* evaluation set. Our approach successfully leverages the pose diversity contained in real-world images to improve performance from the 'Canonical' initialization point close to the version trained on synthetic images (which contain far less inherent real-world diversity and hence the signal is easier to extract). Importantly, 'Ours (real)' outperforms large-scale supervised models, despite using orders of magnitude less 3D supervision.

Fig. 3 demonstrates our scalability to other object categories, *i.e.,* chimpanzees and elephants, trained using real images. The image-to-3D foundation models struggle with complex articulations, frequently hallucinating limbs or producing overly straight, 'canonicalized' shapes due to strong symmetry priors. In contrast, our template-fitting approach guarantees anatomical correctness, and performs better on challenging, asymmetric articulations.

4.3 Ablations

In Tab. 3 we ablate the main components of our method on the *Horse-Mixed* test split. Our starting point is a *Canonical (base)* model, trained exclusively on rest-pose renders with standard Gaussian-sampled camera viewpoints (see the supplementary material for details). Replacing this with heavy-tailed Student's t camera sampling immediately yields substantial gains by exposing the initialization network to extreme, outlier viewpoints. Next, we introduce our iterative pipeline (*+ Refinement w/o reg.*), which extracts posed training data from images. While this provides a substantial performance leap, greedy point-cloud fitting produces structural artifacts. Adding geometric penalties (*+ Regularization*) resolves this by enforcing physically plausible articulations. Finally, *+ Viewpoint augmentation* augments each fitted mesh with diverse novel viewpoints during rendering, forming our complete method. Fig. 4 qualitatively compares the performance of our full model trained on real-world images to the canonical initialization and fully supervised DualPM.

We additionally report a *Random poses baseline*, which trains the model on stochastically deformed meshes with unconstrained, unnatural articulations. Surprisingly, its relatively strong performance highlights that raw pose diversity alone provides a strong training signal (see supplementary material for details).

Fig. 5 presents an ablation on the number of iterations used to train our approach. We observe a large jump in performance across all metrics after the first iteration of our method when compared to the canonical baseline. This improvement continues, approaching the performance of DualPM, but eventually plateaus.

Table 3: Ablation study on the Horse-Mixed test split. Quantitative results evaluating the impact of removing individual pipeline components. Lower scores are better.

Ablation variant	mCD (cm)	% (mCD>5cm)	$\Delta R(\circ)$
DualPM (ground-truth 3D)	3.04	4.00	2.21
Random poses baseline	5.22	44.5	5.12
Canonical (base)	10.04	92.0	12.23
+ Student's t camera sampling	7.24	68.4	7.51
+ Refinement (w/o fitting reg.)	5.59	55.1	6.73
+ Regularization in mesh fitting	4.85	35.0	5.89
+ Viewpoint augmentation ('Ours')	3.86	11.3	3.49

4.4 Limitations

While significantly reducing the amount of 3D supervision needed compared to DualPM, our approach still needs some limited 3D data during training. Specifically, we require a rigged and skinned mesh in a canonical rest pose. However, in practice this is not difficult to acquire for common object categories and automated rigging methods could be used if only a mesh is available. As demonstrated by our experimental results, our approach successfully extracts pose diversity from real-world image collections to improve the predictor. However, Tab. 2 shows that a performance gap remains between training on real-world images, synthetic images, and ground-truth 3D.

5 Conclusion

Supervised learning-based approaches for 3D object reconstruction are only as effective as the data they are trained on. Unlike 2D images, which are available in abundance on the internet, diverse and high quality 3D datasets are much more challenging to create. To sidestep this issue, we introduced a new scalable and automated approach for 3D data generation called BAT3R. Starting from only a single rigged mesh and 2D image collection as input, we showed that it is possible to bootstrap from this initial limited data, via an iterative generation approach, to create diverse and articulated 3D training data. We evaluated the effectiveness of our approach on the challenging task of articulated 3D object reconstruction from a single input image. We approach the performance of fully supervised methods across multiple categories, even when only provided with a fraction of the ground-truth 3D training data.

Acknowledgements. JZ was supported by the UKRI (grant EP/S023208/1), UKRI Centre for Doctoral Training in Robotics and Autonomous Systems at the University of Edinburgh, School of Informatics. OMA was supported by a Royal Society Research Grant. HB was supported by the EPSRC Visual AI grant EP/T028572/1.

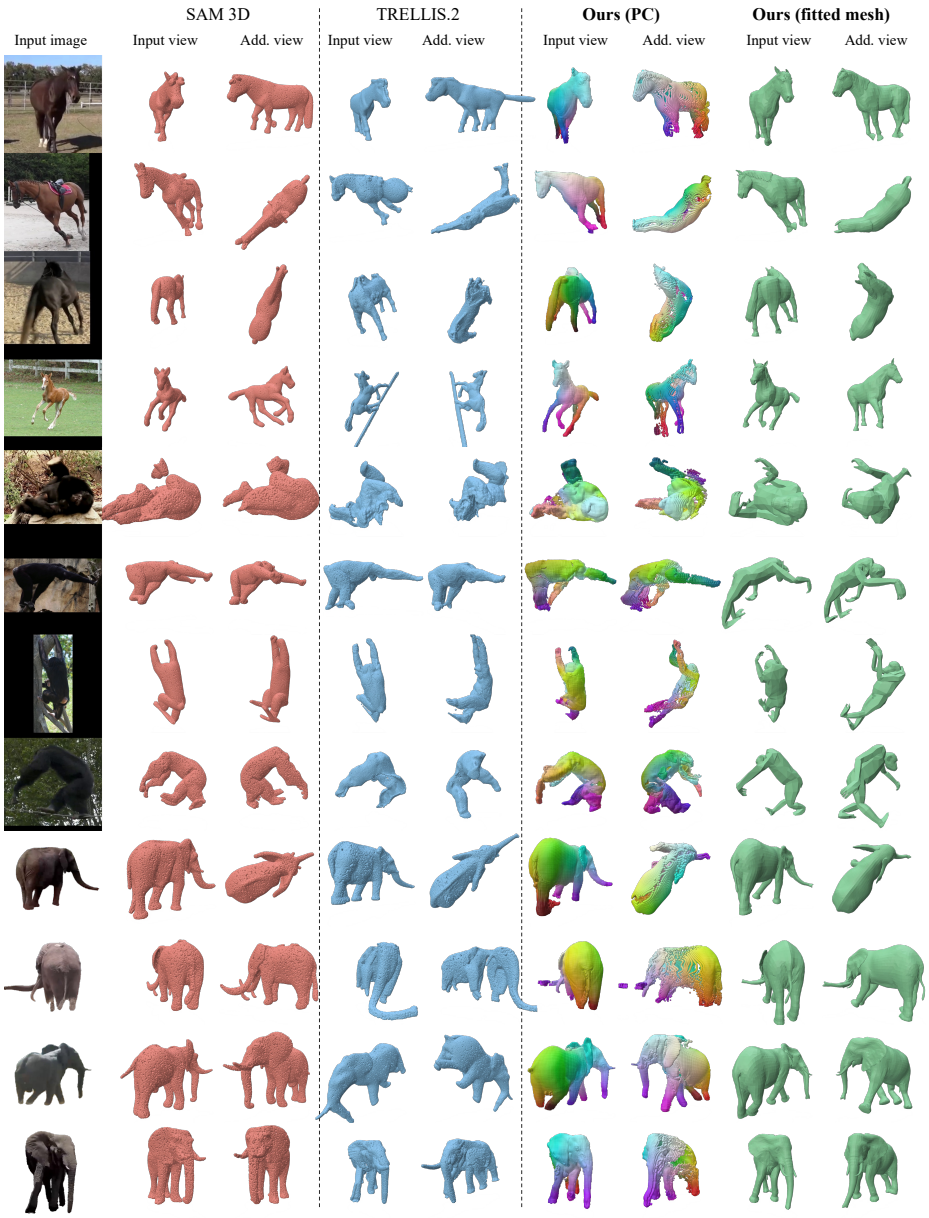

Fig. 3: Qualitative comparison on real-world images across three challenging categories: horse, chimpanzee, and elephant. Trellis.2 [38] struggles with accurate real-world conditioning and relies on post-hoc rigid ICP alignment of their predicted 3D shape to a reference predicted point cloud (here we use 'Ours'). SAM-3D [4] exhibits strong generalization but struggles with complex articulations. Strong symmetry priors often cause hallucinated/missing limbs, overly straight 'canonicalized' shapes, and failures on extreme out-of-distribution poses (*e.g.,* the prone chimpanzee). In contrast, our predicted point clouds capture smooth deformations with natural curvature while disambiguating symmetric parts. Moreover, our template-fitting guarantees anatomical correctness and recovers challenging articulations.

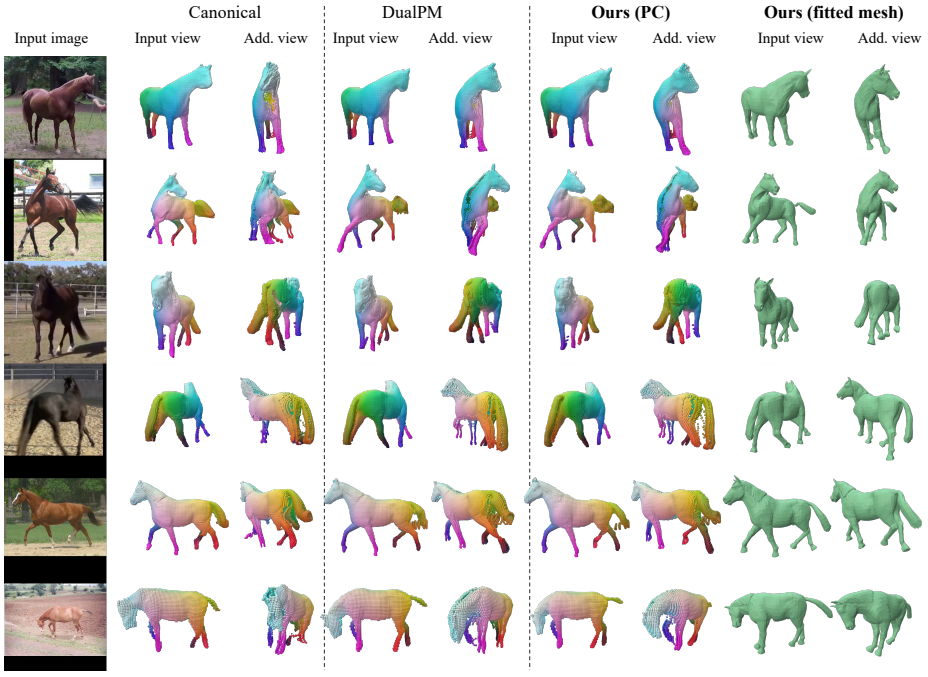

Fig. 4: Qualitative comparison on real-world horse images. We visualize point cloud predictions (input and novel views) and our extracted fitted meshes. While the 'Canonical' baseline captures the coarse pose, it struggles with complex geometry. It exhibits noisy topology, missing parts, and mixed-up appendages (*e.g.,* confusing the rear legs). In contrast, our full model resolves these artifacts, yielding clean geometry that closely matches the fully supervised DualPM.

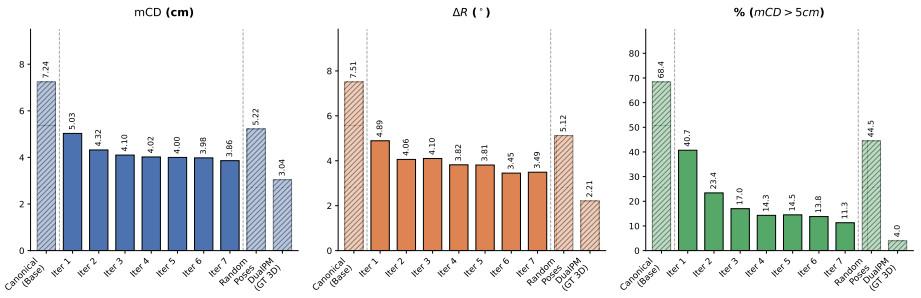

Fig. 5: Performance across refinement iterations. Errors drops significantly after the first iteration compared to the canonical baseline, then improves over successive iterations and plateaus while approaching the fully supervised DualPM (rightmost bars). Notably, the outlier metric % (mCD>5cm) continues to show consistent improvements up through the seventh iteration.

References

1. Aygun, M., Mac Aodha, O.: Saor: Single-view articulated object reconstruction. In: CVPR (2024)
2. Black, M.J., Patel, P., Tesch, J., Yang, J.: Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: CVPR (2023)
3. Cai, Z., Zhang, M., Ren, J., Wei, C., Ren, D., Lin, Z., Zhao, H., Yang, L., Loy, C.C., Liu, Z.: Playing for 3d human recovery. TPAMI (2024)
4. Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., et al.: Sam 3d: 3dfy anything in images. In: CVPR (2026)
5. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. NeurIPS (2023)
6. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: CVPR (2023)
7. Desai, N., Bala, P., Richardson, R., Raper, J., Zimmermann, J., Hayden, B.: Openapepose: a database of annotated ape photographs for pose estimation. eLife (2023)
8. Goel, S., Kanazawa, A., Malik, J.: Shape and viewpoint without keypoints. In: ECCV (2020)
9. Jakab, T., Li, R., Wu, S., Rupprecht, C., Vedaldi, A.: Farm3d: Learning articulated 3d animals by distilling 2d diffusion. In: CVPR (2024)
10. Kabsch, W.: A solution for the best rotation to relate two sets of vectors. Foundations of Crystallography (1976)
11. Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J.: Learning category-specific mesh reconstruction from image collections. In: ECCV (2018)
12. Kaye, B., Jakab, T., Wu, S., Ruprecht, C., Vedaldi, A.: Dualpm: dual posed-canonical point maps for 3d shape and pose reconstruction. In: CVPR (2025)
13. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM TOG (2023)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV (2023)
15. Kulkarni, N., Gupta, A., Fouhey, D.F., Tulsiani, S.: Articulation-aware canonical surface mapping. In: CVPR (2020)
16. Kulkarni, N., Gupta, A., Tulsiani, S.: Canonical surface mapping via geometric cycle consistency. In: ICCV (2019)
17. Li, X., Liu, S., Kim, K., De Mello, S., Jampani, V., Yang, M.H., Kautz, J.: Self-supervised single-view 3d reconstruction via semantic consistency. In: ECCV (2020)
18. Li, Y., Takehara, H., Taketomi, T., Zheng, B., Nießner, M.: 4dcomplete: Non-rigid motion estimation beyond the observable surface. In: ICCV (2021)
19. Li, Z., Litvak, D., Li, R., Zhang, Y., Jakab, T., Rupprecht, C., Wu, S., Vedaldi, A., Wu, J.: Learning the 3d fauna of the web. In: CVPR (2024)
20. Liu, I., Xu, Z., Yifan, W., Tan, H., Xu, Z., Wang, X., Su, H., Shi, Z.: Riganything: Template-free autoregressive rigging for diverse 3d assets. ACM TOG (2025)
21. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: ICCV (2023)
22. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: CVPR (2024)

23. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries (2023)
24. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM (2021)
25. Monnier, T., Fisher, M., Efros, A.A., Aubry, M.: Share with thy neighbors: Single-view reconstruction by cross-instance consistency. In: ECCV (2022)
26. Murez, Z., Van As, T., Bartolozzi, J., Sinha, A., Badrinarayanan, V., Rabinovich, A.: Atlas: End-to-end 3d scene reconstruction from posed images. In: ECCV (2020)
27. Niewiadomski, T., Yiannakidis, A., Cuevas-Velasquez, H., Sanyal, S., Black, M.J., Zuffi, S., Kulits, P.: Generative zoo. In: ICCV (2025)
28. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: ICLR (2025)
29. Rueegg, N., Zuffi, S., Schindler, K., Black, M.J.: Barc: Breed-augmented regression using classification for 3d dog reconstruction from images. IJCV (2023)
30. Rüegg, N., Tripathi, S., Schindler, K., Black, M.J., Zuffi, S.: Bite: Beyond priors for improved three-d dog pose estimation. In: CVPR (2023)
31. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
32. Song, C., Li, X., Yang, F., Xu, Z., Wei, J., Liu, F., Feng, J., Lin, G., Zhang, J.: Puppeteer: Rig and animate your 3d models. In: NeurIPS (2025)
33. Tesch, J., Becherini, G., Achar, P., Yiannakidis, A., Kocabas, M., Patel, P., Black, M.J.: Bedlam2. 0: Synthetic humans and cameras in motion. In: NeurIPS (2025)
34. Vicente, S., Carreira, J., Agapito, L., Batista, J.: Reconstructing pascal voc. In: CVPR (2014)
35. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024)
36. Wu, S., Jakab, T., Rupprecht, C., Vedaldi, A.: Dove: Learning deformable 3d objects by watching videos. IJCV (2023)
37. Wu, S., Li, R., Jakab, T., Rupprecht, C., Vedaldi, A.: Magicpony: Learning articulated 3d animals in the wild. In: CVPR (2023)
38. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., et al.: Native and compact structured latents for 3d generation. In: CVPR (2026)
39. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: CVPR (2025)
40. Xu, J., Zhang, Y., Peng, J., Ma, W., Jesslen, A., Ji, P., Hu, Q., Zhang, J., Liu, Q., Wang, J., et al.: Animal3d: A comprehensive dataset of 3d animal pose and shape. In: ICCV (2023)
41. Yang, Z., Cai, Z., Mei, H., Liu, S., Chen, Z., Xiao, W., Wei, Y., Qing, Z., Wei, C., Dai, B., et al.: Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In: ICCV (2023)
42. Yao, C.H., Hung, W.C., Li, Y., Rubinstein, M., Yang, M.H., Jampani, V.: Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. NeurIPS (2022)
43. Yao, C.H., Hung, W.C., Li, Y., Rubinstein, M., Yang, M.H., Jampani, V.: Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In: CVPR (2023)

44. Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D.: Ap-10k: A benchmark for animal pose estimation in the wild. In: NeurIPS Track on Datasets and Benchmarks (2021)
45. Zhang, J.P., Pu, C.F., Guo, M.H., Cao, Y.P., Hu, S.M.: One model to rig them all: Diverse skeleton rigging with unirig. ACM TOG (2025)
46. Zhao, B.N., Wu, J., Wu, S.: Web-scale collection of video data for 4d animal reconstruction. In: NeurIPS Datasets and Benchmarks Track (2026)
47. Zou, Z.X., Yu, Z., Guo, Y.C., Li, Y., Liang, D., Cao, Y.P., Zhang, S.H.: Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In: CVPR (2024)
48. Zuffi, S., Kanazawa, A., Jacobs, D.W., Black, M.J.: 3d menagerie: Modeling the 3d shape and pose of animals. In: CVPR (2017)
49. Zuffi, S., Mellbin, Y., Li, C., Hoeschle, M., Kjellström, H., Polikovsky, S., Hernlund, E., Black, M.J.: Varen: Very accurate and realistic equine network. In: CVPR (2024)