# Generative Zoo

Tomasz Niewiadomski[1] Anastasios Yiannakidis[1] Hanz Cuevas-Velasquez[1] Soubhik Sanyal[1]
Michael J. Black[1] Silvia Zuffi[2] Peter Kulits[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]IMATI-CNR, Milan, Italy
{tomasz,ayiannakidis,hcuevas,ssanyal,black,kulits}@tue.mpg.de, silvia.zuffi@cnr.it

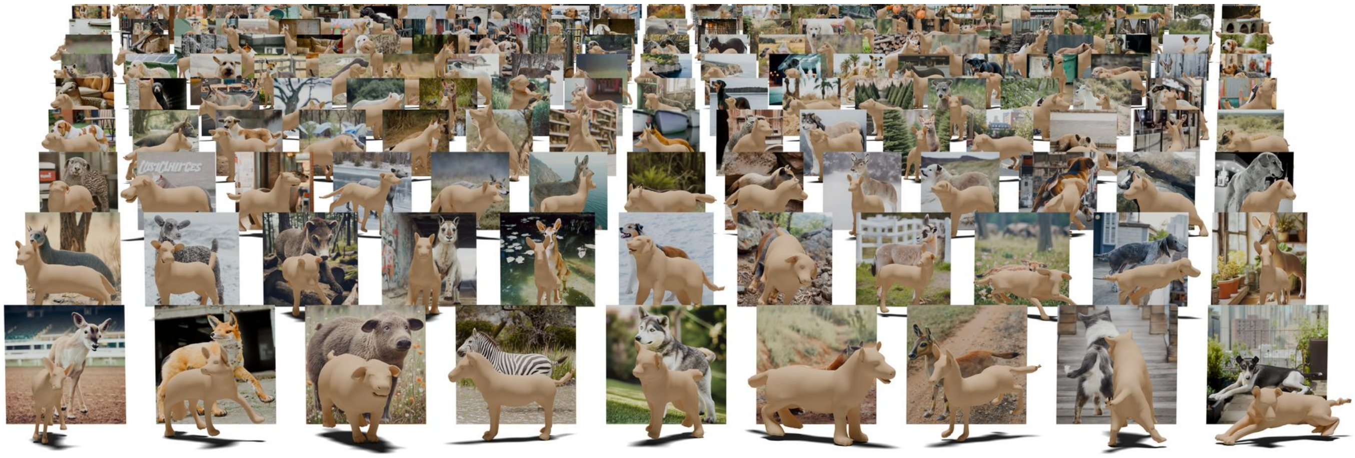

Figure 1. We propose a pipeline for the scalable generation of realistic 3D animal pose and shape estimation training data. Training solely on samples produced using our pipeline (see pairs above), we achieve state-of-the-art performance on a real-world multi-species benchmark.

## Abstract

*The model-based estimation of 3D animal pose and shape from images enables computational modeling of animal behavior. Training models for this purpose requires large amounts of labeled image data with precise pose and shape annotations. However, capturing such data requires the use of multi-view or marker-based motion-capture systems, which are impractical to adapt to wild animals in situ and impossible to scale across a comprehensive set of animal species. Some have attempted to address the challenge of procuring training data by pseudo-labeling individual real-world images through manual 2D annotation, followed by 3D-parameter optimization to those labels. While this approach may produce silhouette-aligned samples, the obtained pose and shape parameters are often implausible due to the ill-posed nature of the monocular fitting problem. Sidestepping real-world ambiguity, others have designed complex synthetic-data-generation pipelines leveraging game engines and collections of artist-designed 3D assets. Such engines yield perfect ground-truth labels but are often lacking in visual realism and require considerable manual effort to adapt to new species or environments. We propose an alternative approach to synthetic-data generation: rendering with a conditional image-generation model. We introduce a pipeline that samples a diverse set of poses and shapes for a variety of mammalian quadrupeds and generates realistic images with corresponding ground-truth pose and shape parameters. To demonstrate the scalability of our approach, we introduce GenZoo, a synthetic dataset containing one million images of distinct subjects. We train a 3D pose and shape regressor on GenZoo, which achieves state-of-the-art performance on a real-world multi-species 3D animal pose and shape estimation benchmark, despite being trained solely on synthetic data. We release our data and pipeline at https://genzoo.is.tue.mpg.de.*

## 1. Introduction

The estimation of animal pose from images enables the computational modeling of animal behavior [1]. In quantifying behavior, pose offers a low-dimensional representation amenable to analysis [39]. From pose, actions can be segmented [34] or individual health can be monitored [8].

Pose by itself can be a fairly descriptive feature when measured in a laboratory setting, where cameras, lighting, and environmental conditions can be tightly controlled. However, pose becomes less informative in the wild, where environmental conditions can vary greatly. Inspired by trends in the modeling of humans [28, 31], recent work has expanded beyond primitive pose-based representations and toward parametric models that represent not only 3D pose, but also shape [3, 7, 26, 64, 67]. These representations, typically derived from collections of 3D scans, are often steered by body-joint rotation parameters and a latent shape code.

However, estimating the parameters of such representations from images is challenging due to the ill-posed nature of 3D inference from 2D images, the diversity of real-world animals, and the variability of imaging conditions in natural environments. Training an effective regression model typically requires large amounts of annotated data. Datasets curated for human pose and shape estimation, for parametric body models like SMPL [31], include extensive sequences captured using marker-based mocap systems [21] or fitted IMU devices [49]. While humans can be brought into 4D capture halls and outfitted with dense markers, wild animals are less cooperative. As a result, alternative approaches must be devised to overcome the problem of obtaining data.

The most common approach is the manual annotation of 2D landmarks or silhouettes [4, 65], which can be used for weak model supervision. Other approaches begin with 2D annotations and produce pseudo-labels of pose and shape by optimizing a model such as SMAL [64] to fit the labeled 2D features [55]. However, while optimizing to conform to a silhouette may yield 3D fits that appear plausibly aligned, producing accurate 3D annotations is difficult without sufficiently strong priors, as many physically implausible pose and shape combinations might explain the same silhouette.

Recently, the use of video-game engines to produce rendered synthetic data has been explored as an alternative to labeling real-world images [6, 18, 20]. While the approach sacrifices visual realism, it offers greater control over dataset curation. Graphics engines' explicit representation of 3D scenes enables the production of precise ground-truth annotations and control over dataset statistics. However, traditional graphics-based synthetic datasets require substantial manual effort to design or modify. To produce data for additional species, or render them in a new environment, requires a new set of 3D assets. While the data can be made to appear realistic, achieving both visual realism and sufficient diversity requires considerable resources.

We investigate a potentially simpler alternative to the production of synthetic data: *rendering* with a conditional image-generation model. We propose a pipeline that, given the name of a species, produces paired images and ground-truth pose-and-shape parameters. Rather than relying on explicit collections of 3D assets and scenes, our pipeline is controllable by language: the inclusion of a new species or environmental setting is accomplished via prompting. Our pipeline facilitates the generation of realistic images with a degree of control comparable to traditional synthetic-data generators, thus combining the advantages of visual realism, scalability, and the ability to control data production.

To demonstrate the scalability of our approach, we introduce *GenZoo*, a million-image dataset comprised of unique poses and shapes across diverse mammalian quadrupeds; see Fig. 1 for samples. Training a regression model solely on our synthetic dataset, without the use of annotated real-world images, we achieve state-of-the-art performance on Animal3D [55], a real-world multi-species animal pose and shape estimation benchmark, validating the quality of our dataset. We also introduce a synthetic evaluation dataset with greater annotation fidelity than previous benchmarks. In summary, our key contributions include:

1. A scalable pipeline for the generation of synthetic 3D quadrupedal animal pose and shape estimation data
2. *GenZoo*, a million-scale multi-species synthetic dataset
3. A state-of-the-art pose and shape regression model
4. *GenZoo-Felidae*, a high-fidelity synthetic benchmark.

## 2. Related Work

**Animal Pose and Shape Estimation.** The 3D reconstruction of animals follows two primary paradigms: model-free and model-based. Model-free approaches make minimal assumptions about the animal's 3D body structure, and the objective is to obtain a representative 3D surface. Given the diversity observed across animal species and shape, this approach is fairly common. Notable examples include CMR [23], which deformed a spherical mesh to reconstruct birds from images, and LASSIE [61], MagicPony [52], and 3D-Fauna [29], which learned articulated 3D shape from image collections. ViSER [58], LASR [57], BANMo [59], and PPR [60] recovered 3D shape of animals from video.

Model-based approaches alternatively assume that a 3D model is provided (or retrievable [54]), either as a species-specific template model or as a parametric 3D model that captures shape variations within and across species. This approach is particularly valuable for downstream analysis, as 3D pose and shape parameters can be leveraged to estimate and track conformation and behavior over time. Among model-based approaches, Cashman and Fitzgibbon [9] were the first to address the representation challenge, creating a morphable model for dolphins. Later, Kanazawa et al. [22] estimated a deformable 3D model for cats. Zuffi et al. [64] introduced SMAL, an articulated shape model for a variety of quadrupeds learned from scans of toys. SMAL has been adopted to estimate shape and pose for zebras [66], to estimate the 3D pose and shape of dogs [4, 5, 43] from 2D datasets with keypoints and silhouette annotations, and to create dog avatars [45]. Further modeling developments

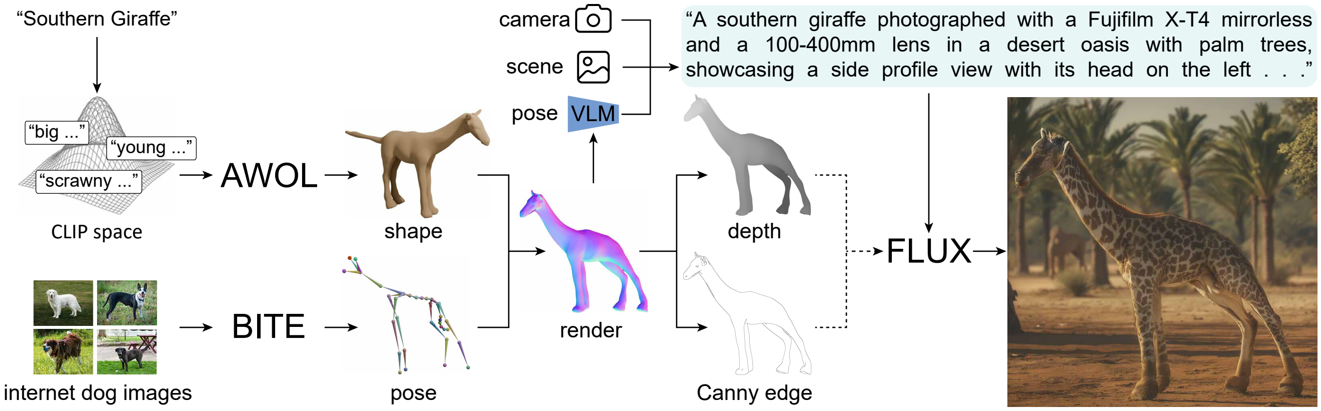
"Southern Giraffe"
"big ..."
"young ..."
"scrawny ..."
CLIP space
AWOL
shape
internet dog images
BITE
pose
camera
scene
pose
VLM
render
"A southern giraffe photographed with a Fujifilm X-T4 mirrorless and a 100-400mm lens in a desert oasis with palm trees, showcasing a side profile view with its head on the left . . ."
depth
Canny edge
FLUX

Figure 2. **Pipeline Overview.** Starting with a sampled animal name (Sec. 3.2), we sample corresponding shape parameters (Sec. 3.3). Paired pose parameters are sampled from a set of pseudo-poses (Sec. 3.4). Sampled camera and scene descriptions are combined with a pose caption to form a prompt (Sec. 3.5). Rendered control signals and the prompt are used to guide the conditional image-generation model, resulting in the final image (Sec. 3.6).

include Rüegg et al. [44]'s BITE extension of SMAL for dog breeds, Li et al. [26]'s use of a horse-specific model for lameness detection, Wang et al. [50]'s learning of a bird model from images, and Zuffi et al. [67]'s pioneering horse model learned from real 4D scans. Zuffi and Black [63] introduced SMAL+ in AWOL, an enhanced version of the SMAL model learned from additional 3D scans. In RAW, Kulits et al. [24] extended the animal-reconstruction problem to additionally model the surrounding environment.

**Rendered Synthetic Data for Pose Estimation.** A number of works approach the generation of synthetic data for human pose and shape estimation. SURREAL [48] sampled random backgrounds and applied cloth textures to posed SMPL [31] meshes. AGORA [37] rendered images of clothed-body scans with SMPL-X [38] ground-truth annotations. BEDLAM [6] extended this, presenting a dataset with simulated clothing, hair, and large variation in human shape. Hewitt et al. [18, 19] applied displacement maps on a modified SMPL body to simulate natural cloth wrinkles.

In contrast, synthetic animal datasets lack comparable sophistication. This is due in part to the lack of extensive motion datasets like AMASS [33], but also to the greater morphological variation between animals. Several methods have employed mesh renders to train a 2D joint regressor for a single species, including for mice [7], cougars [14], and dogs [46]. However, none of these methods can be easily combined to train a multi-animal 3D pose and shape regressor. Although Mu et al. [36] produced a dataset of more than ten different animals, it can only be applied to learn 2D joint estimation and not 3D pose and shape, as it lacks variation in shape. Li et al. [27] introduced PFERD, a marker-based horse motion-capture dataset obtained with dense motion capture, including a diverse set of body shapes and sizes.

**Generative Models and Training Data.** Recent advances in controllable image generation, such as Stable Diffusion [42] and ControlNet [62], enable realistic data production for downstream tasks [2, 32, 51]. DatasetDM [53] finetuned Stable Diffusion to generate synthetic images and ground-truth pairs for depth and human pose estimation as well as semantic and instance segmentation. Others have focused on human pose and shape estimation data, to synthesize new samples [15, 51, 62] or augment existing ones [10].

## 3. Method

In this section, we present our approach for generating synthetic training data for 3D animal pose and shape estimation (see Fig. 2). After introducing the SMAL body model (Sec. 3.1), which defines our pose and shape representation, we introduce our pipeline. Starting with a set of mammalian species or breeds (Sec. 3.2), we sample a taxon. Based on the taxon, we sample an animal shape (Sec. 3.3) and assign a pose (Sec. 3.4). From the model parameters, we render a primitive image, which is captioned using a vision–language model (VLM) and used to synthesize a prompt (Sec. 3.5). Finally, we condition an image-generation model using both the prompt and render (Sec. 3.6). We close with an explanation of our regression-model baseline (Sec. 3.7).

### 3.1. SMAL

The *Skinned Multi-Animal Linear* (SMAL) [64] model is a function that, given shape parameters $\beta$ and pose parameters $\theta$, transforms a 3D template to produce a posed mesh. The transformation occurs in two steps: first, the vertex template instance $\mathbf{v}_t$ is deformed into an intrinsic shape $\mathbf{v}_s$, then Linear Blend Skinning (LBS) is applied to rotate the deformed body parts based on the pose parameters $\theta$:

$$\begin{aligned}\mathbf{v}_s &= \mathbf{v}_t + B\beta^T \\ \mathbf{v} &= \mathrm{LBS}(\mathbf{v}_s, \theta; W, J_r).\end{aligned} \quad (1)$$

Here, the template $\mathbf{v}_t$ represents the initial state of a triangular mesh with $n_V$ vertices, $B$ is a matrix of shape $3n_V \times n_B$ containing the $n_B$ basis vectors of a linear shape deformation space, $J_r$ is the joint regressor that maps model vertices to a set of $n_J$ 3D joint locations, and $W$ is a skinning weight matrix used in LBS. The linear shape space is learned using Principal Component Analysis (PCA) on a set of scans of toy quadrupedal animals. In particular, we employ SMAL+, an expanded variant of SMAL introduced in AWOL [63], learned from a set of 145 registered toy scans, including figures covering a number of mammalian species.

## 3.2. Species Sampling

Our approach offers a key advantage over traditional rendering-based synthetic-data generation: rather than requiring additional artist-designed 3D assets, adding a new species requires only prompt modifications. This enables precise control over dataset sampling statistics: new taxon can be readily added or species proportions rebalanced.

To maximize data diversity, we sample from a variety of taxa listed in the Mammal Diversity Database [11]. While SMAL can represent a wide range of quadrupedal animals, its fixed joint topology and limitated shape-space expressivity constrain the set of representable taxa. As a result, we restrict our sampling to mammals within the superorder Laurasiatheria (e.g. giraffes, deer, cows), excluding members of the order Eulipotyphla (e.g. rats, moles). See Fig. 3 in the Supp. Mat. visualizing the taxonomy we sample from.

We note that breeds are considered distinct from taxonomical species. For example, *Canis familiaris* (dog) is regarded as a single species. Considering the diversity in shape, size, and visual appearance across breeds of dogs, we expand the single dog category with a set of 247 breeds when sampling. We sample dogs with 50% probability. Once species or breed has been sampled, we generate a 3D, posed animal and a corresponding image through steps described in Secs. 3.3 to 3.6 and laid out visually in Fig. 2.

## 3.3. Shape Sampling

AWOL [63] is a recent flow-based generative model that maps CLIP [40] embeddings to SMAL shape parameters, producing shapes conditioned on an embedding of an animal text description. While generating shapes well-aligned with a given prompt, AWOL is non-stochastic: each input maps to one shape, making it difficult to cover the space of possible shapes. Avoiding AWOL altogether and sampling betas naively would result in implausible shapes. Representativeness and diversity are competing objectives. Instead, we sample in CLIP embedding space and use AWOL for decoding. For each taxon, we compute 128 CLIP text embeddings, using a list of appearance descriptors such as “big,” “young,” or “scrawny.” CLIP is prompted with “A photo of a X Y.” We fit a multivariate Gaussian distribution to the resulting CLIP embeddings, from which we can then sample to generate stochastic shapes that maintain alignment with the taxon. This enables non-naive sampling of (species-aligned) shape with the deterministic AWOL, introducing controlled variability through sampling of the distribution.

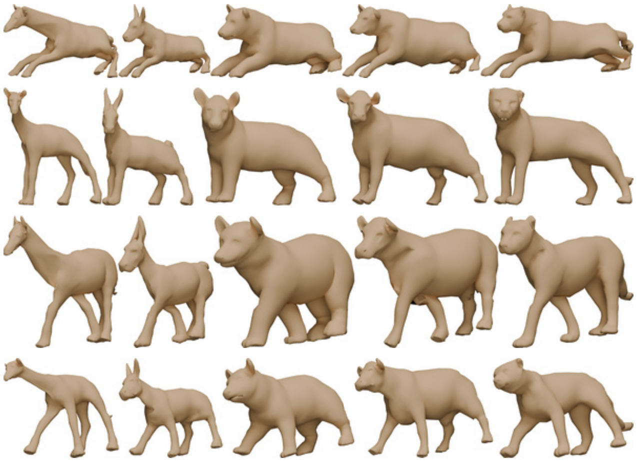

Figure 3. **Pose Transfer.** The same set of pseudo poses transferred across individuals with different body-shape / limb proportions.

## 3.4. Pose Sampling

A similar dilemma between diversity and realism arises in the sampling of ground-truth pose. To maximize diversity, one could sample random rotations. This is feasible for human body models like SMPL [31] where vast amounts of motion-capture data are available, which can be used to learn a generative pose prior [38]. However, the absence of comparable data for animals or SMAL makes sampling valid poses more difficult and requires different strategies.

To address this data limitation, we apply BITE [44], an optimization-based dog-pose estimation method, to process a large collection of online dog images and extract plausible pseudo-poses. Although BITE does not extract accurate poses from every image, we use only the extracted pose, discarding the original photos. We then sample from this collection of poses to populate our generated dataset.

We source our poses from dogs because of: 1) the large number of dog photos available online, 2) their dynamic nature and relative flexibility, and 3) the broad range of body shapes and proportions “dog” encompasses, which lead to a variety of different poses. In collecting pseudo-poses, our goal is not to match exactly the joint positions of the animal in the source image, but to cover well the space of possible quadruped poses when sampling new ones. Not all poses we employ are realistic on every body shape. For example, one might not expect to encounter in the wild a jumping cow. However, it’s possible, and we hope to represent it in our dataset. See Fig. 3 for a visualization of a set of poses transferred across individuals of varying body proportions.

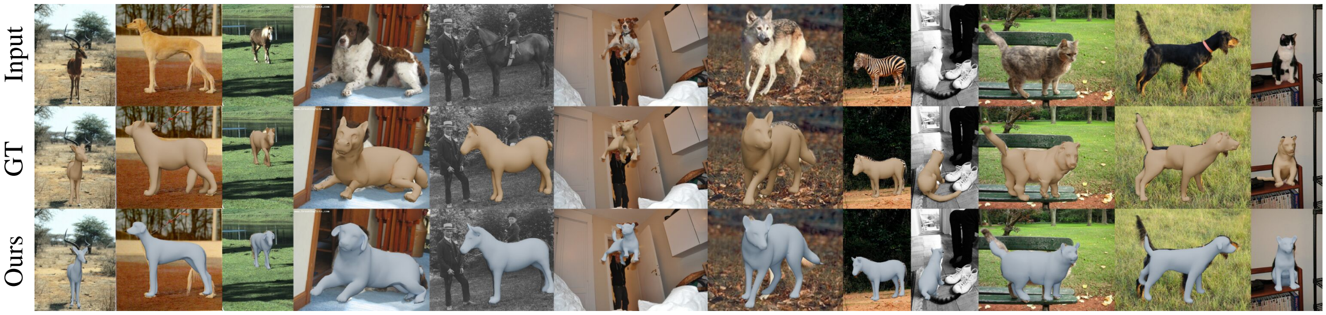

Figure 4. **Animal3D Reconstruction Samples.** We show the input image (top), GT mesh (middle), and our model's prediction (bottom).

### 3.5. Prompt Sampling

After sampling species, shape, and pose, we render out the resulting model in Pyrender [35]. The render is passed to a VLM, Molmo-7B-D-0924 [12], prompted with "This is a picture of an animal. Which direction is the animal facing?" The caption provides guidance on global orientation, helping to reduce ambiguity in the subsequent image-generation step, particularly for producing images of complex poses.

In addition to captioning, and inclusion of the species name, we also sample a camera setting (e.g., "Samsung Galaxy S23 with enhanced night mode camera") and a scenery setting (e.g., "bike trail through the woods") from predefined lists. We find that explicit prompting of these attributes increases visual realism and diversity. After the above are sampled, we pass the descriptors to an LLM, Qwen2.5-7B-Instruct [47], to synthesize these components into a coherent and concise prompt. See Sec. 4.4 for a quantitative ablation study on our prompting design decisions.

### 3.6. Conditional Image Generation

Once the prompt has been prepared, we employ FLUX [25], a generative text-to-image rectified-flow transformer model, to synthesize paired images. While FLUX can be conditioned on CLIP [40] and T5 [41] embeddings, it cannot natively be controlled with pose and shape parameters. To achieve such control, we employ an auxiliary, pretrained ControlNet [62] model and perform no finetuning of it.

ControlNets are task-specific generators of a generative model – FLUX in our case – trained to maximize output probability conditioned on provided control signals, such as a Canny-edge or depth map. Using Pyrender, we produce a depth map and a shaded render for Canny-edge extraction. During generation, FLUX and the ControlNet are used concurrently to guide the generation process by both the text prompt and extracted control signals. We generate all images at a resolution of 1024x1024. See Fig. 2 for a visual of the control signals applied, Sec. 4.3 for an ablation on the control signal used, and also Supp. Mat. Sec. C for an ablation on the choice of image-generation model.

### 3.7. Parameter Regressor

We train a regression model on *GenZoo* using two architectures. Following the results of an ablation study performed by Goel et al. [16], we build off the backbone of ViT-Pose [56] and add a parameter-estimation head. This backbone was originally trained for human 2D-keypoint estimation. To match the baselines employed in Animal3D [55], we additionally train our model with ResNet-50 [17]. We supervise the training of the model using losses on 2D-joint re-projection and directly on pose and parameters. See Supp. Mat. Sec. A for additional details on reproducibility.

## 4. Evaluations

We quantitatively evaluate the regressor trained on our data using three metrics on joint positions: 1) PCK@0.5, defined as the percentage of correct keypoints within half the head–tail length; 2) PA-MPJPE, the procrustes-aligned mean per-joint positional error in mm; and 3) S-MPJPE, defined as PA-MPJPE without the rotation transform in mm, as used in Animal3D [55] to account for SMAL scale inconsistency.

### 4.1. Animal3D

We employ Animal3D [55] to evaluate the transferability of our model trained on *GenZoo*. Animal3D is built on a set of images borrowed from the ImageNet [13] and COCO [30] datasets. Annotators manually labeled 2D keypoints and silhouettes, which were used to guide an optimization-based fitting process, resulting in SMAL pseudo-labels. It contains animals of forty classes, which correspond to ImageNet labels of a subset of the superorder Laurasiatheria.

Training solely on *GenZoo*, we achieve state-of-the-art performance, outperforming the strongest baseline by 57% in S-MPJPE (374.9→160.1), see Tab. 1. We also observe notable improvements in PCK@0.5 (85.6→97.0) and PA-MPJPE (123.9→116.6), but gains appear comparatively saturated. See Fig. 4 for reconstruction samples, Fig. 5 for a qualitative comparison, and Fig. 6 for performance by class. PCK@0.5, which we report in order to compare with

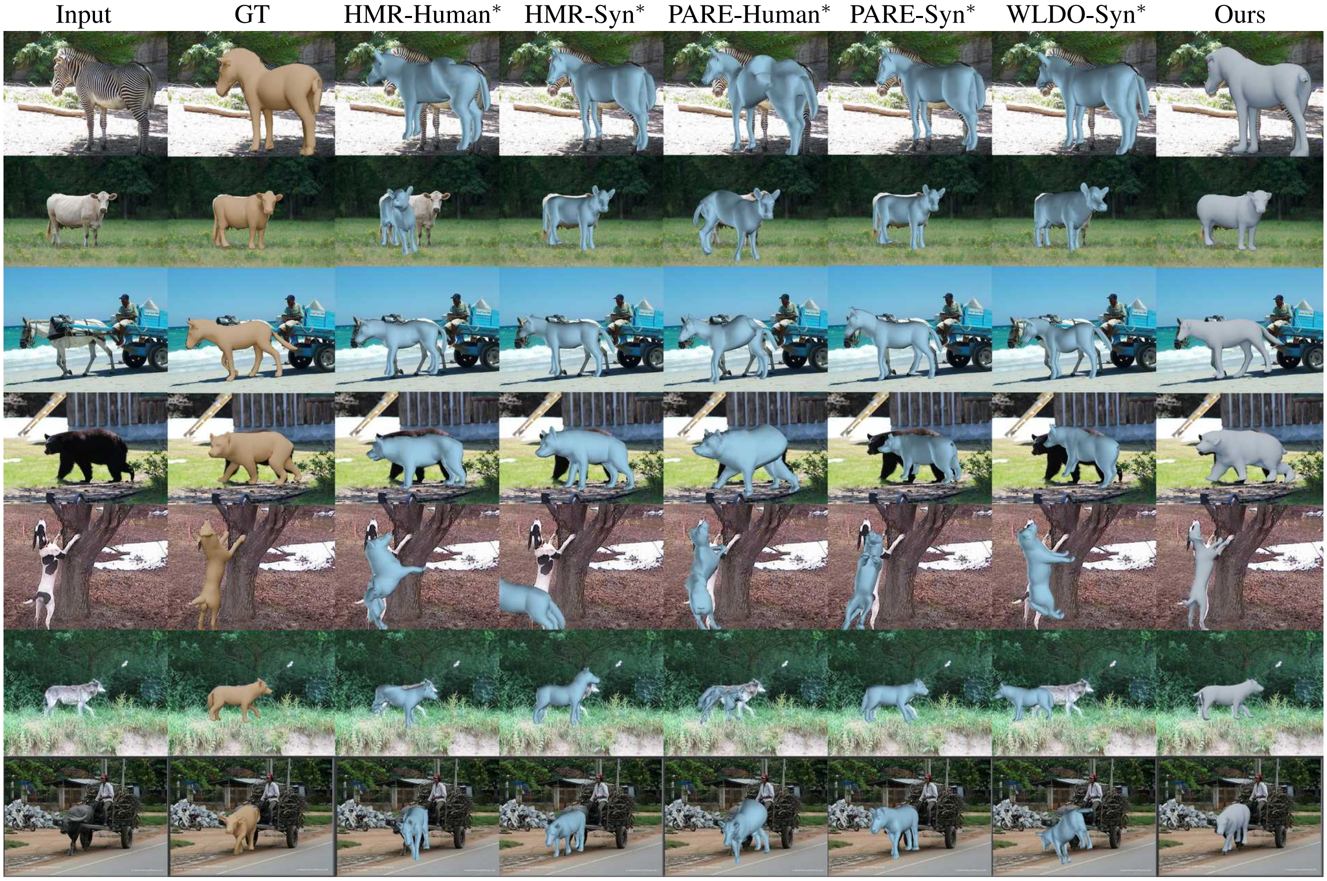

Figure 5. **Qualitative Method Comparison.** Predictions between our method and the baseline results (*) sourced from Animal3D [55].

| | ↑ PCK@0.5 | ↓ S-MPJPE | ↓ PA-MPJPE |
|---|---|---|---|
| Ours | **97.0** | **160.1** | **116.6** |
| Ours (ResNet) | <u>95.11</u> | <u>201.1</u> | 132.67 |
| HMR* | 63.1 | 496.2 | 124.8 |
| PARE* | 85.6 | 374.9 | 127.2 |
| WLDO* | 65.1 | 484.0 | <u>123.9</u> |

Table 1. **Quantitative Method Comparison**. We compare our models trained on *GenZoo* with the best-performing numbers on the Animal3D benchmark. Asterisks (*) signify the result is from Animal3D and the best number was chosen across experiments.

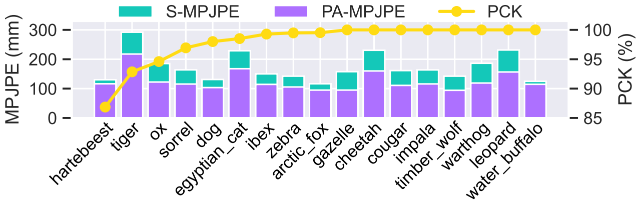

Figure 6. **Animal3D Performance By Species.** Our model outperforms the SOTA aggregate of 374.9 S-MPJPE across species, demonstrating the cross-species generalization of our model.

the numbers reported in Animal3D, is a very generous metric. Despite the large improvement in S-MPJPE, we observe only modest difference over the baselines in PA-MPJPE.

To investigate this, we look to the ground-truth. We visualize in Fig. 7 an Animal3D annotation next to our model prediction. While we observe that the projection of the ground-truth is well-aligned with the image silhouette, when viewed from the side it appears as a very different animal, highlighting the difficulty of producing accurate manual annotations without sufficient priors. This suggests there may be a bound on performance. We explore this further in a perceptual study detailed in Supp. Mat. Sec. B.

### 4.2. GenZoo-Felidae

Motivated by our observations in Sec. 4.1, we propose a complementary benchmark. We design it to both evaluate model generalization to unseen species and to test estimation ability against ground-truth shape. We use the *GenZoo* data-generation pipeline to create a 1000-sample test set

| | Animal3D | | | GenZoo-Felidae | | | | |
|---|---|---|---|---|---|---|---|---|
| | ↑ PCK@0.5 | ↓ S-MPJPE | ↓ PA-MPJPE | ↑ PCK@0.5 | ↓ S-MPJPE | ↓ PA-MPJPE | ↓ S-V2V | ↓ PA-V2V |
| Full | 97.1 | **166.9** | **118.4** | **99.6** | 83.5 | 62.0 | 91.5 | 72.1 |
| -Depth | 96.7 | 184.1 | 135.1 | 99.1 | 114.2 | 83.3 | 131.1 | 101.8 |
| -Canny | 96.2 | 172.3 | 119.4 | 99.5 | **81.0** | **59.7** | **87.5** | **68.7** |
| -Caption | 96.9 | 167.1 | 120.1 | 99.5 | 93.5 | 71.0 | 104.3 | 82.0 |
| -LLM | **97.2** | 168.2 | 120.7 | 99.5 | 92.4 | 70.3 | 101.3 | 81.2 |

Table 2. **Quantitative Ablation Effects.** Ablation-study results for models trained on 100,000 samples each.

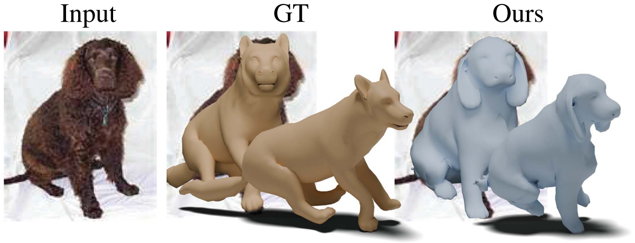

Figure 7. **Animal3D Comparison.** Highlighting the difficulty of producing manual 3D annotations of monocular, real-world images, we observe physical implausibilities in the Animal3D ground-truth. In contrast, a model trained on our dataset does not exhibit the same biases. See Supp. Mat. Sec. B for a perceptual study comparing our predictions with the dataset's pseudolabels.

with no overlap in species, pose, camera setting, or environmental conditions. We extract individual poses from various artist-designed feline motion sequences retargeted to the SMAL skeleton. We designate the 47 species in the Felidae family as test-only and do not train on them. We refer to our benchmark as *GenZoo-Felidae*. We show dataset samples and reconstructions in Fig. 10 and metrics in Tab. 2.

### 4.3. Control Signal

We evaluate different ControlNet conditioning combinations, comparing our full model (combined Canny-edge and depth control) against the controls applied individually. Depth-only conditioning produces the most-realistic images but shows poorest ground-truth alignment. In contrast, Canny-edge-only conditioning achieves the best alignment but compromises visual realism. We observe that combining the two at reduced strength balances these tradeoffs. See Fig. 8 for an illustrative sample. We find that quantitative evaluation supports our qualitative findings (Tab. 2).

### 4.4. Captioning and Textual Conditioning

We assess the impact of removing the VLM captioning step and LLM prompt synthesis. While both components positively contribute to Animal3D performance metrics, their impact is most pronounced in *GenZoo-Felidae*. See Tab. 2.

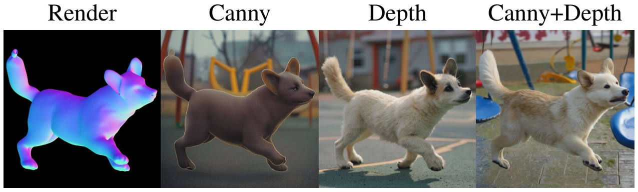

Figure 8. **Control-Signal Ablation.** Depth-only conditioning produces the most realistic images but poorest alignment with ground-truth poses. Canny-edge-only control shows the opposite effect. We employ both to balance visual realism with pose accuracy.

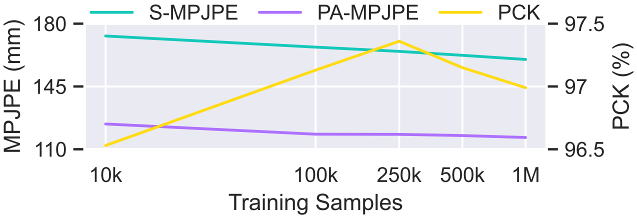

Figure 9. **Data Efficiency.** We evaluate data efficiency of our model and observe consistent log-linear trends as data is scaled.

### 4.5. Data Efficiency

We evaluate our model trained with varying amounts of data. While we observe a positive quantitative trend as the amount of training data increases, the returns appear diminishing. This further suggests that there may be an upper bound on Animal3D performance. See Fig. 9 for a plot.

## 5. Discussion and Limitations

While our *GenZoo*-trained model demonstrates fairly robust generalization to real-world images, several limitations warrant discussion: 1) Our model struggles under strong occlusions, such as mistaking a human in the foreground as the regression target. See Fig. 11 for examples. See also `https://genzoo.is.tue.mpg.de` for a comprehensive set of reconstruction visualizations and Fig. 2 in the Supp. Mat. for reconstructions beyond Animal3D images. See also Supp. Mat. Sec. E, for an ablation study on

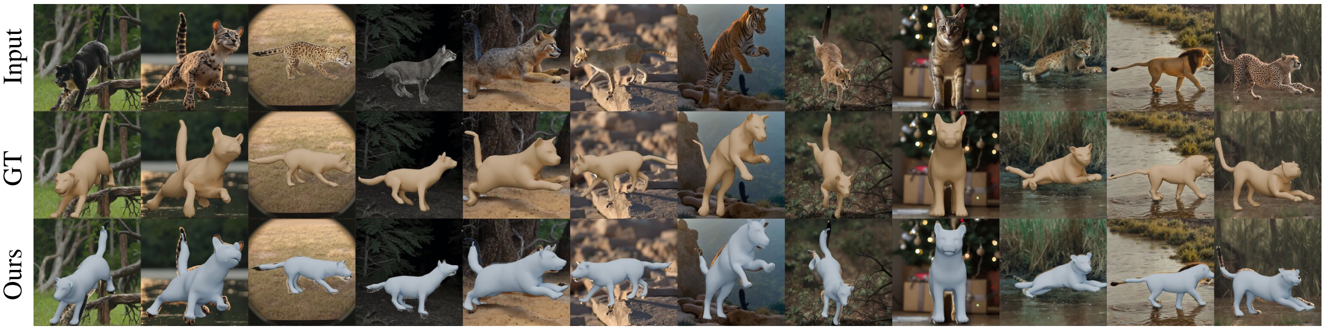

Figure 10. **GenZoo-Felidae Dataset Samples.** We show the input image (top), GT mesh (middle), and our model's prediction (bottom).

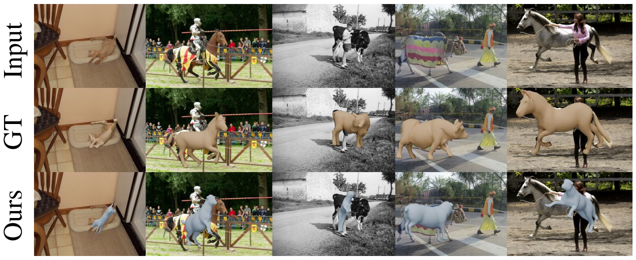

Figure 11. **Prediction Failures.** Our model can struggle when faced with strong occlusion or uncommon species-specific poses.

augmenting with occlusions during training to improve the model's robustness. 2) While our pose-sampling distribution, built from dog pseudo poses, appears sufficient for a broad coverage of animal poses, the model can struggle with species-specific poses not typically observed in dogs, such as feline grooming positions. Future work should explore more-comprehensive pose-sampling strategies. 3) Although the SMAL model is capable of representing well a broad swathe of shapes and species, it is constrained by its fixed skeletal topology and cannot represent large morphological differences between species such as the trunk of an elephant. Future work should focus on developing or utilizing more-expressive parametric representations that can accommodate greater anatomical diversity, or finer-grained species-specific representations such as VAREN [67]. 4) We observe that FLUX has limited understanding relating to lesser-known species, often instead producing a more common, taxonomically similar species, but sometimes mistaking the name for a tropical bird (Fig. 12). Future work should explore the adaptation of image-generation models to better represent rare and unusual species. 5) While there is a lack of training data for 3D animal pose-and-shape estimation, there is also a need for strong benchmarks with precise ground-truth annotations. *GenZoo* – as well as *GenZoo-Felidae* – while comparable, do not offer the degree of precision of traditional synthetic data. This includes ambiguity in control, such as relative-depth conditioning.

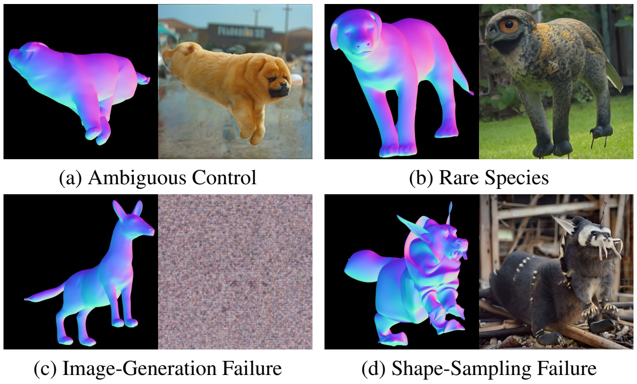
(a) Ambiguous Control
(b) Rare Species
(c) Image-Generation Failure
(d) Shape-Sampling Failure

Figure 12. **Generation Failures.** Generation failure cases most often arise from: (a) ambiguous control signals; (b) uncommon species names mistaken for exotic birds (Formosan serow); (c) image-generation model failure; and (d) shape-sampling errors. Such failures might be automatically filtered through 2D and 3D consistency checks between generated images and ground truth.

## 6. Conclusion

Motivated by the shortcomings of existing approaches for the acquisition of 3D animal pose and shape estimation training data, we proposed a scalable pipeline that leverages conditional image-generation models. Our pipeline enables the generation of realistic images with a degree of control comparable to that of traditional synthetic-data generators. Showcasing the scalability of our approach, we presented *GenZoo*, a dataset of one million images of unique animals. Training solely on *GenZoo*, without the use of any real-world training data, we demonstrated state-of-the-art performance on a real-world multi-species animal pose and shape estimation benchmark. We additionally introduced *GenZoo-Felidae*, a high-fidelity synthetic test dataset that complements existing pseudo-labeled real-world evaluations. Beyond immediate technical achievements, our work opens new possibilities for automated animal behavior analysis, wildlife monitoring, and veterinary applications.

**Acknowledgments.** We thank Tsvetelina Alexiadis and Taylor McConnell for perceptual study counsel, Arina Kuznetcova and Nikos Athanasiou for moral support. **Disclosure.** MJB has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. MJB has financial interests in Amazon and Meshcapade GmbH. While MJB is a co-founder and Chief Scientist at Meshcapade, his research in this project was performed solely at, and funded solely by, the Max Planck Society. SZ is supported by PNRR FAIR Future AI Research (PE00000013), Spoke 8 Pervasive AI (CUP H97G22000210007) and NBFC National Biodiversity Future Center (CN00000033), Spoke 4 (CUP B83C22002930006) under the NRRP MUR program by NextGenerationEU.

## References

[1] David J. Anderson and Pietro Perona. Toward a science of computational ethology. *Neuron*, 84(1):18–31, 2014. 1

[2] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves ImageNet classification. *TMLR*, 2023. 3

[3] Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd G. Pfrommer, Marc F. Schmidt, and Kostas Daniilidis. 3D bird reconstruction: A dataset, model, and shape recovery from a single view. In *ECCV*, pages 1–17, Berlin, Heidelberg, 2020. Springer-Verlag. 2

[4] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and SMAL: Recovering the shape and motion of animals from video. In *ACCV*, pages 3–19, Cham, 2019. Springer International Publishing. 2

[5] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3D animal reconstruction with expectation maximization in the loop. In *ECCV*, pages 195–211. Springer, 2020. 2

[6] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, 2023. 2, 3

[7] Luis A. Bolaños, Dongsheng Xiao, Nancy L. Ford, Jeff M. LeDue, Pankaj K. Gupta, Carlos Doebeli, Hao Hu, Helge Rhodin, and Timothy H. Murphy. A three-dimensional virtual mouse generates synthetic training data for behavioral analysis. *Nature Methods*, 18(4):378–381, 2021. 2, 3

[8] Sofia Broomé, Marcelo Feighelstein, Anna Zamansky, Gabriel Carreira Lencioni, Pia Haubro Andersen, Francisca Pessanha, Marwa Mahmoud, Hedvig Kjellström, and Albert Ali Salah. Going deeper than tracking: A survey of computer-vision based recognition of animal pain and emotions. *IJCV*, 131(2):572–590, 2023. 1

[9] Thomas J. Cashman and Andrew W. Fitzgibbon. What shape are dolphins? Building 3D morphable models from 2D images. *IEEE TPAMI*, 35(1):232–244, 2013. 2

[10] Hanz Cuevas-Velasquez, Priyanka Patel, Haiwen Feng, and Michael J. Black. Toward human understanding with controllable synthesis, 2024. 3

[11] Mammal Diversity Database. Mammal diversity database, 2024. 4

[12] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, pages 91–104, 2025. 5

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5

[14] Abassin Sourou Fangbemi, Yi Fei Lu, Maoyuan Xu, Xiaowu Luo, Alexis Rolland, and Chedy Raissi. ZooBuilder: 2D and 3D pose estimation for quadrupeds using synthetic data. In *Eurographics/ ACM SIGGRAPH Symposium on Computer Animation - Showcases*. The Eurographics Association, 2020. 3

[15] Yongtao Ge, Wenjia Wang, Yongfan Chen, Yang Liu, Hao Chen, Xuan Wang, and Chunhua Shen. 3D human reconstruction in the wild with synthetic data using generative models, 2024. 3

[16] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, pages 14783–14794, 2023. 5

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[18] Charlie Hewitt, Tadas Baltrušaitis, Erroll Wood, Lohit Petikam, Louis Florentin, and Hanz Cuevas-Velasquez. Procedural humans for computer vision, 2023. 2, 3

[19] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrušaitis. Look ma, no markers: Holistic performance capture without the hassle. *ACM TOG*, 43(6), 2024. 3

[20] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G. Schwing. SAIL-VOS: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In *CVPR*, pages 3105–3115, 2019. 2

[21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE TPAMI*, 36(7):1325–1339, 2014. 2

[22] Angjoo Kanazawa, Shahar Kovalsky, Ronen Basri, and David Jacobs. Learning 3D deformation of animals from

2D images. In *Eurographics*, pages 365–374, Goslar, DEU, 2016. Eurographics Association. 2

[23] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, pages 371–386, 2018. 2

[24] Peter Kulits, Michael J. Black, and Silvia Zuffi. Reconstructing animals and the wild. In *CVPR*, pages 16565–16577, 2025. 3

[25] Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2022. 5

[26] Ci Li, Nima Ghorbani, Sofia Broomé, Maheen Rashid, Michael J. Black, Elin Hernlund, Hedvig Kjellström, and Silvia Zuffi. hSMAL: Detailed horse shape and pose reconstruction for motion pattern recognition, 2021. 2, 3

[27] Ci Li, Ylva Mellbin, Johanna Krogager, Senya Polikovsky, Martin Holmberg, Nima Ghorbani, Michael J. Black, Hedvig Kjellström, Silvia Zuffi, and Elin Hernlund. The poses for equine research dataset (PFERD). *Scientific Data*, 11(497): 1, 2024. 3

[28] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM TOG*, 36(6), 2017. 2

[29] Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3D fauna of the web. In *CVPR*, pages 9752–9762, 2024. 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, Cham, 2014. Springer International Publishing. 5

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 2, 3, 4

[32] Wufei Ma, Qihao Liu, Jiahao Wang, Angtian Wang, Xiaoding Yuan, Yi Zhang, Zihao Xiao, Guofeng Zhang, Beijia Lu, Ruxiao Duan, Yongrui Qi, Adam Kortylewski, Yaoyao Liu, and Alan Yuille. Generating images with 3D annotations using diffusion models. In *ICLR*, 2024. 3

[33] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019. 3

[34] Jesse D. Marshall, Tianqing Li, Joshua H. Wu, and Timothy W. Dunn. Leaving flatland: Advances in 3D behavioral measurement. *Current Opinion in Neurobiology*, 73:102522, 2022. 1

[35] Matthew Matl. Pyrender. `https://github.com/mmatl/pyrender`, 2022. 5

[36] Jiteng Mu, Weichao Qiu, Gregory D. Hager, and Alan L. Yuille. Learning from synthetic animals. In *CVPR*, pages 12386–12395, 2020. 3

[37] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, pages 13468–13478, 2021. 3

[38] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 3, 4

[39] Talmo D. Pereira, Joshua W. Shaevitz, and Mala Murthy. Quantifying behavior to understand the brain. *Nature Neuroscience*, 23(12):1537–1549, 2020. 1

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 4, 5

[41] Adam Roberts, Hyung Won Chung, Gaurav Mishra, Anselm Levskaya, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Kehang Han, Michelle Casbon, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. Scaling up models and data with t5x and seqio. *JMLR*, 24(377):1–8, 2023. 5

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3

[43] Nadine Rüegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. BARC: Learning to regress 3D dog shape from images by exploiting breed information. In *CVPR*, pages 3876–3884, 2022. 2

[44] Nadine Rüegg, Shashank Tripathi, Konrad Schindler, Michael J. Black, and Silvia Zuffi. BITE: Beyond priors for improved three-D dog pose estimation. In *CVPR*, pages 8867–8876, 2023. 3, 4

[45] Remy Sabathier, Niloy J. Mitra, and David Novotny. Animal avatars: Reconstructing animatable 3D animals from casual videos. In *ECCV*, pages 270–287. Springer Nature Switzerland, 2025. 2

[46] Moira Shooter, Charles Malleson, and Adrian Hilton. SyDog-Video: A synthetic dog video dataset for temporal pose estimation. *IJCV*, 132(6):1986–2002, 2024. 3

[47] Qwen Team. Qwen2.5: A party of foundation models, 2024. 5

[48] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, pages 109–117, 2017. 3

[49] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *ECCV*, pages 601–617, 2018. 2

[50] Yufu Wang, Nikos Kolotouros, Kostas Daniilidis, and Marc Badger. Birds of a feather: Capturing avian shape models from images. In *CVPR*, pages 14739–14749, 2021. 3

[51] Zhenzhen Weng, Laura Bravo-Sánchez, and Serena Yeung-Levy. Diffusion-HPC: Synthetic data generation for human mesh recovery in challenging domains. In *3DV*, pages 257–267. IEEE, 2024. 3

[52] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. MagicPony: Learning articulated 3D animals in the wild. In *CVPR*, pages 8792–8802, 2023. 2

[53] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. DatasetDM: Synthesizing data with perception annotations using diffusion models. *NeurIPS*, 36:54683–54695, 2023. 3

[54] Yuefan Wu, Zeyuan Chen, Shaowei Liu, Zhongzheng Ren, and Shenlong Wang. CASA: Category-agnostic skeletal animal reconstruction. In *NeurIPS*, pages 28559–28574. Curran Associates, Inc., 2022. 2

[55] Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, et al. Animal3D: A comprehensive dataset of 3D animal pose and shape. In *ICCV*, pages 9099–9109, 2023. 2, 5, 6

[56] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, pages 38571–38584. Curran Associates, Inc., 2022. 5

[57] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T. Freeman, and Ce Liu. LASR: Learning articulated shape reconstruction from a monocular video. In *CVPR*, pages 15980–15989, 2021. 2

[58] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. ViSER: Video-specific surface embeddings for articulated 3D shape reconstruction. In *NeurIPS*, pages 19326–19338. Curran Associates, Inc., 2021. 2

[59] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. BANMo: Building animatable 3D neural models from many casual videos. In *CVPR*, pages 2853–2863, 2022. 2

[60] Gengshan Yang, Shuo Yang, John Z. Zhang, Zachary Manchester, and Deva Ramanan. PPR: Physically plausible reconstruction from monocular videos. In *ICCV*, pages 3914–3924, 2023. 2

[61] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. LASSIE: Learning articulated shapes from sparse image ensemble via 3D part discovery. In *NeurIPS*, pages 15296–15308. Curran Associates, Inc., 2022. 2

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3, 5

[63] Silvia Zuffi and Michael J. Black. AWOL: Analysis without synthesis using language. In *ECCV*, pages 1–19, Cham, 2025. Springer Nature Switzerland. 3, 4

[64] Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*, pages 6365–6373, 2017. 2, 3

[65] Silvia Zuffi, Angjoo Kanazawa, and Michael J. Black. Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images. In *CVPR*, pages 3955–3963, 2018. 2

[66] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J. Black. Three-D safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *ICCV*, pages 5359–5368, 2019. 2

[67] Silvia Zuffi, Ylva Mellbin, Ci Li, Markus Hoeschle, Hedvig Kjellström, Senya Polikovsky, Elin Hernlund, and Michael J. Black. VAREN: Very accurate and realistic equine network. In *CVPR*, pages 5374–5383, 2024. 2, 3, 8