WildAni4D: Towards 4D Animal Mesh Reconstruction

Gyeongsu Cho[1] Hezhen Hu[2] Donghyeon Soon[3] Changwoo Kang[1] Kyungdon Joo[1*]
[1]UNIST [2]University of Texas at Austin [3]DGIST
{threedv,kangchangwoo,kyungdon}@unist.ac.kr
alexhu@utexas.edu dhsoon@dgist.ac.kr

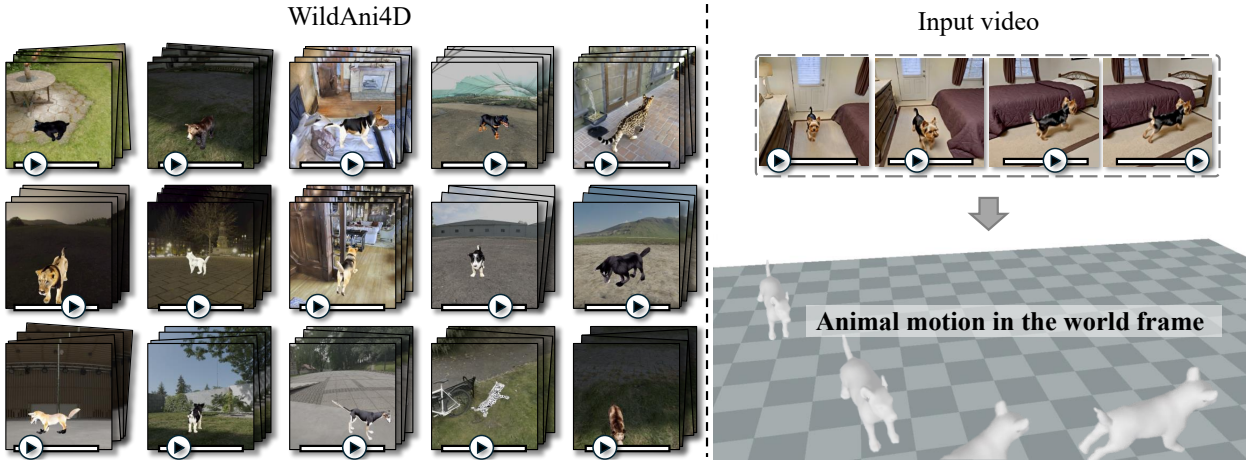

Figure 1. **WildAni4D: A Synthetic Video Generation Pipeline and Reconstruction Model for 4D Animal Motion.** Left: Samples from our generator: SMAL-based animals with sequence-consistent shape, realistic textures, and diverse indoor/outdoor scenes and camera motions; each frame is annotated with SMAL parameters, 2D/3D keypoints, and segmentation. Right: 4D Reconstruction of Animal from in-the-wild. Given an in-the-wild video, our model reconstructs metric-scale global trajectory.

Abstract

Recovering 4D animal motion, including 3D geometry and global trajectory, is essential for quantitative biomechanics and behavioral analysis. However, two major challenges hinder the progress of animal motion recovery. Existing methods lack sufficient annotated video data and suffer from per-frame temporal instability. To address this, we introduce WildAni4D, a framework that unites a novel synthetic video generation pipeline with a new reconstruction model. First, our generator creates a large-scale dataset of realistic, appearance-consistent video sequences with ground-truth animal geometry and camera trajectories. Second, our reconstruction model robustly estimates temporally coherent motion using a single sequence-level shape and per-frame pose predictions. We demonstrate that our model outperforms state-of-the-art per-frame methods, drastically reducing temporal pose flicker and shape drift. WildAni4D offers a scalable solution for 4D animal reconstruction, enabling large-scale motion analysis from in-the-wild videos. Moreover, WildAni4D enables diverse downstream applications, including animal motion data annotation, animatable animal reconstruction, and text-to-motion generation. Project page: vision3d-lab.github.io/wildani4d

1. Introduction

Accurately recovering the 4D motion of animals is essential for understanding animal behavior, building animatable digital animals, and supporting biomechanical analysis. However, unlike humans, animals exhibit extreme diversity in appearance, scale, and motion patterns, and handheld monocular cameras often capture animals in unconstrained environments. As a result, reconstructing world-grounded 4D animal motion from in-the-wild videos remains an unsolved challenge.

* Corresponding author.

Compounding the problem is the lack of 4D animal training data. In the human domain, large motion capture datasets [11, 22] and multi-view recordings have enabled strong video models and world-grounded motion estimation [28, 35]. For animals, however, collecting pose-annotated, trajectory-aligned, multi-view videos at scale is effectively impossible. Consequently, most existing animal approaches focus on single-image 3D recovery [21, 23, 27], which cannot handle temporal dynamics, suffer from shape drift over time, and cannot recover global trajectories.

In this paper, we address these challenges by introducing a complete system for recovering the global trajectory and articulated motion of 3D animals from monocular in-the-wild videos. First, we develop a large-scale synthetic video generation pipeline. We combine dynamic, textured 3D animals, which are generated with Skinned Multi-Animal Linear Model (SMAL) [47] mesh from text-prompted shapes [46] and motion dataset [17, 41], with diverse 3D scenes. These scenes are sourced from indoor scans (*e.g.*, Matterport [5]), outdoor 3DGS captures [10, 14], and HDRI maps. We render these elements with a variety of complex, simulated camera trajectories to mimic real-world videography (see Fig. 1).

Second, we introduce the Animal Video Transformer (AVT), a two-stage reconstruction model built to explicitly disentangle camera motion from animal motion. Using this camera path as a global reference, AVT regresses the animal's kinematic motion. We add a shape consistency module by modifying the regression head to output a single shape for the entire sequence, while predicting pose per frame. This design enforces shape consistency across the video and suppresses the severe temporal drift that affects per-frame or naïve sequential models. By composing the recovered global camera motion with the regressed relative animal motion, our method achieves accurate and temporally coherent 4D animal reconstruction in the world space.

We demonstrate that trained on our synthetic dataset, AVT achieves the strongest performance among the compared methods on our synthetic benchmark, reducing pose error and improving temporal coherence relative to single-frame baselines. On in-the-wild videos, it also yields more faithful results under a downstream reconstruction-based evaluation protocol. Moreover, we demonstrate that our framework enables diverse downstream applications, including animal motion data annotation, animatable animal reconstruction, and text-to-motion generation.

In summary, our contributions are:

- We propose a synthetic video generation pipeline that combines dynamic textured animals, diverse 3D scenes, and realistic camera motions to produce realistic, annotated training data.
- We propose the Animal Video Transformer (AVT), a complete two-stage system that recovers global 4D animal motion from monocular video by integrating spatio-temporal features with a novel shape consistency regression scheme for robust and stable reconstruction.
- We highlight that the proposed framework can support diverse downstream applications, including pseudo-ground-truth annotation, animatable animal reconstruction, and text-to-motion generation.

2. Related Work

Animal Pose and Shape Reconstruction. The 3D reconstruction of animals is a critical task for the quantitative analysis of behavior and biomechanics. This field faces significant challenges not present in human-centric domains, including vast species shape diversity and frequent joint occlusions inherent to quadrupedal locomotion. The 3D reconstruction of animals is divided into model-free approaches, which reconstruct 3D shape directly from images, and model-based approaches. Model-free methods, such as LASSIE [42], MagicPony [36], 3DFauna [18], BANMO [38] and RAC [39], learn to reconstruct articulated 3D shapes without a predefined template, offering broad applicability. However, these representations are not always explicitly editable or suited for kinematic animation.

In contrast, model-based approaches leverage a parametric 3D template. The seminal work in this area is the SMAL model [47], a parametric model for quadrupeds learned from 3D scans of toy figures. This model has been widely adopted and extended [6, 16, 26], enabling methods to estimate 3D pose and shape from monocular in-the-wild images [3, 26, 27, 48]. Nevertheless, SMAL estimation has not been thoroughly examined in other challenging species, despite the high-quality reconstruction that has been achieved on specific species (*e.g.*, horses and dogs).

To address this problem, recent efforts have focused on scalable pipelines for generating realistic synthetic images. GenZoo [23] and AniMer [21] have introduced pipelines that leverage conditional image-generation models [44] to create large-scale static image datasets with perfect ground-truth SMAL parameters. Both methods adopt HMR2.0 [8]-style per-frame ViT architectures [7, 37], enabling strong single-image pose and shape estimation when trained on sufficiently large synthetic datasets. These datasets have proven highly effective for training robust, single-frame 3D animal reconstruction models. However, animal behavior is an inherently dynamic process. To bridge this gap, we propose a 4D animal reconstruction model that maintains strong temporal consistency across frames.

Synthetic Data for Pose Estimation. The challenge of data acquisition is not unique to the animal domain. In the 3D Human Mesh Recovery (HMR) task, synthetic data has become a cornerstone for training robust models [4, 24, 40]. The BEDLAM dataset [4], in particular, demonstrated a

critical breakthrough. By generating a large-scale synthetic video dataset with high-fidelity graphics and realistic, physics-simulated clothing, BEDLAM showed that a network trained only on synthetic data could achieve state-of-the-art accuracy on real-world image benchmarks [13, 34].

A key insight from BEDLAM is that data quality and scale can be as important as architectural innovation. BEDLAM showed that high-quality synthetic data can substantially advance HMR. This suggests that the primary bottleneck is often the data, not the model. Earlier synthetic animal datasets also support the effectiveness of synthetic supervision. SyDog [30], SyDog-Video [32], and Digidogs [31] showed that synthetic data can improve 2D, temporal, and single-view 3D dog pose estimation, respectively, although these works focus on dog-specific pose estimation rather than general 4D animal mesh reconstruction.

We hypothesize that this principle holds for the animal domain. While recent generative pipelines have improved the realism of static images [21, 23], they lack the dynamic and environmental complexity required to learn video-based motion reconstruction. Our work is analogous to BEDLAM, but for animals. We introduce a high-fidelity video-centric pipeline designed to provide the necessary data to train robust 3D animal motion estimation models that generalize to real-world videos.

3D Human Model Recovery from Videos. Our goal of recovering full 3D animal motion from a monocular video is directly informed by extensive research in the human domain [8, 12, 35]. The fundamental challenge is to jointly recover both the kinematic body motion and the subject's global trajectory in the world space, especially when the camera is also moving [43]. Recent human-centric methods, such as GVHMR [28] and WHAM [29], tackle this by learning strong motion priors from large-scale Motion Capture datasets [22] to directly regress the human's world-space trajectory. This approach, however, is heavily reliant on the diversity of the training data. In contrast, TRAM proposes a scene-centric, two-stage approach that explicitly disentangles camera and human motion. It first utilizes a robustified monocular SLAM [33] to estimate the metric-scale camera trajectory by relying only on the static scene background. This camera motion is then composed with a temporally-refined local body motion estimate to recover the full, world-grounded human trajectory.

This latter, scene-centric method is particularly compelling for the animal domain. The animal domain lacks the equivalent of large-scale, high-fidelity MoCap datasets. Data scarcity of animals renders the development of robust, prior-based trajectory regressors [28, 29] currently infeasible. We hypothesize that a decoupled approach, inspired by TRAM [35], offers a more viable path forward. By leveraging the static environment as a metric-scale reference frame via DROID-SLAM [33], we can circumvent the need for motion priors learned from non-existent MoCap data. Our work investigates a two-stage strategy to recover the complete 3D animal motion by disentangling the estimation of the camera trajectory from the animal local body motion.

3. Method

Our goal is to recover the 3D animal motion, including its global trajectory and articulated body movement, from monocular in-the-wild videos. Our approach consists of two main components. First, we develop a high-fidelity synthetic data generation pipeline (Sec. 3.2) to produce annotated video data. This pipeline is built upon the Skinned Multi-Animal Linear (SMAL) [47] model (Sec. 3.1). Second, we introduce a video-based transformer model (Sec. 3.3) that learns to reconstruct an animal global trajectory and 3D motion by training on our synthetic data.

3.1. SMAL

Our work uses the SMAL [47], a differentiable parametric model that maps a shape vector β and a pose vector θ to a 3D animal mesh v. The model factorizes this transformation into a two-stage process. First, a mean template v_t is offset by linear blend shapes B according to β to create an unposed, identity-specific shape v_s:

$$v_s = v_t + B\beta^T. \tag{1}$$

Second, this shaped mesh v_s is articulated using standard Linear Blend Skinning (LBS). This stage is driven by the pose parameters θ, a joint regressor J_r that defines joint locations, and a set of skinning weights W:

$$v = \text{LBS}(v_s, \theta; W, J_r). \tag{2}$$

We specifically employ the SMAL+ variant, an enhanced model introduced in AWOL [46]. This variant is learned from registered 3D scans, providing a more expressive 145-dimensional shape space ($\beta \in \mathbb{R}^{145}$).

3.2. Synthetic Video Generation Pipeline

To train our reconstruction model, we generate a large-scale synthetic video dataset. Our pipeline comprises three main components: (1) dynamic textured animals, (2) realistic 3D scenes, and (3) diverse camera motions.

Dynamic Animal Generation. For each T-frame sequence, we first generate a static animal identity. Following [23], we sample a 145-dimensional shape parameter, β, which remains constant for all T frames. This β is generated using the AWOL model from a text prompt (*e.g.*, "a photo of a [adjective] [animal]"). For articulated motion, we sample a T-frame pose sequence $\{\theta_t\}_{t=1}^{T}$ from AnimalML3D [41], where each θ_t denotes the SMAL pose parameters at frame t. To generate a high-fidelity texture,

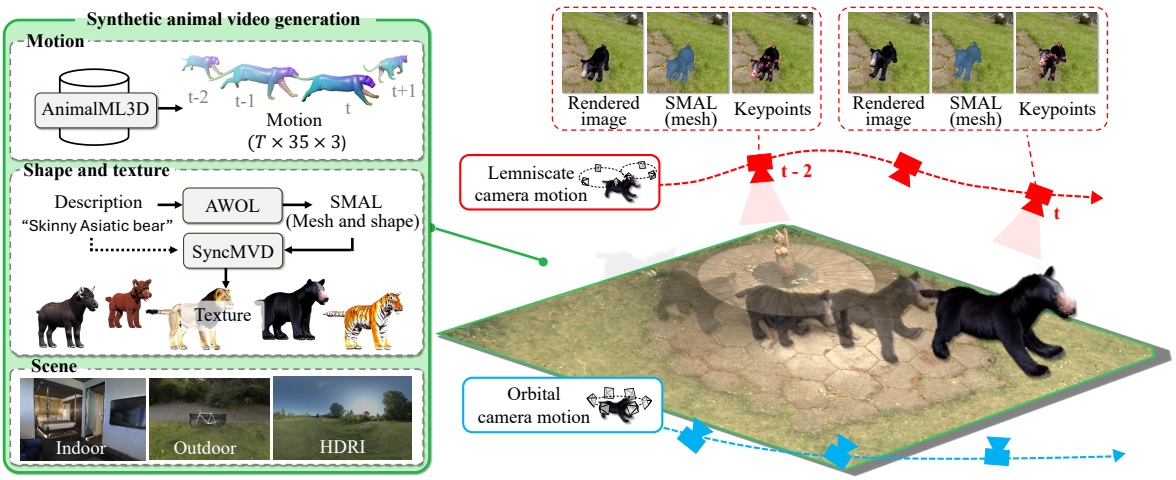

Figure 2. **WildAni4D-Gen: Synthetic Animal Video Generation Pipeline.** Our generation pipeline is a fully 3D rendering system that synthesizes realistic animal videos by combining motion sequences, SMAL-based shape and texture generation, diverse 3D scenes, and simulated camera trajectories to produce fully annotated synthetic data.

we employ a text-to-mesh texture generation model (*e.g.*, SyncMVD [19]). We use the generated SMAL mesh as the 3D template and condition the texture model with the same text prompt used for shape generation, as shown in Fig. 2.

3D Scene Synthesis. We build our environments from three diverse sources to ensure robustness. (1) Indoor Scans: We utilize reconstructed 3D meshes from the Matterport dataset [5]. (2) Outdoor Captures: We leverage 3DGS [14] captures of outdoor scenes from the MIP-NeRF 360 dataset [1], which we convert to textured meshes using a reconstruction method SuGaR [10]. (3) HDRI Domes: For unbounded backgrounds, we project a high-dynamic-range image (HDRI) onto a large 3D dome mesh.

Camera Motion Synthesis. As part of our data generation pipeline, we simulate four distinct camera trajectories to mimic different animal-camera motion relationships observed in real-world videos: (1) Static: The camera position and rotation are fixed. (2) Orbit: The camera translates to approximately follow the movement of the animal while orbiting around it (same elevation). (3) Lemniscate: The camera follows a figure-eight path relative to the animal trajectory. (4) Track-and-Rotate: The camera translation is fixed, but it rotates to keep the animal in frame.

By rendering the dynamic animal within the 3D scene using these camera motions, we generate a complete video with corresponding ground truth 3D annotations, including per-frame SMAL parameters, keypoints, and segmentation masks. We render animal videos using Pytorch3D [25].

3.3. 3D Animal Motion Reconstruction

Following TRAM [35], we propose a two-stage method to recover the animal global trajectory and body motion. First, we estimate the metric-scale global camera motion from the video. Second, we use this camera motion as a reference frame for our Animal Video Transformer (AVT) model, which regresses the animal's kinematic motion.

Global Trajectory Estimation. We first recover the camera trajectory up to an unknown scale using a SLAM algorithm. Specifically, we employ a Masked DROID-SLAM [33] that masks out the dynamic animal, ensuring that only static background features are used for bundle adjustment. To recover the metric scale, we align the relative depth map from SLAM with a metric depth map predicted by a metric depth estimation network (*e.g.*, ZoeDepth [2]). This provides a metric-scale camera trajectory $\{G_t\}_{t=1}^{T}$ in $SE(3)$ which serves as the reference frame.

Animal Video Transformer (AVT). We introduce the Animal Video Transformer (AVT), a transformer-based architecture for regressing 3D animal motion from video (see Fig. 3). Inspired by recent ViT-based mesh recovery models [8, 37], AVT builds on a pre-trained Vision Transformer (ViT) backbone. In contrast to single-frame methods [8, 21, 23], AVT applies temporal modeling after the ViT backbone, not within the decoder. Specifically, each frame is first encoded into patch tokens by the frozen ViT backbone, and a temporal transformer is then applied to the sequence of tokens to learn spatio-temporal features across frames before regression. The resulting temporally enriched tokens are finally passed to the transformer decoder, which predicts the motion parameters.

A key design choice of AVT is to enforce shape consistency. As in our synthetic data generation pipeline (Sec. 3.2), an animal's shape parameter β remains constant within a sequence. Accordingly, the regression head predicts a single sequence-level shape parameter $\hat{\beta}$, while pose

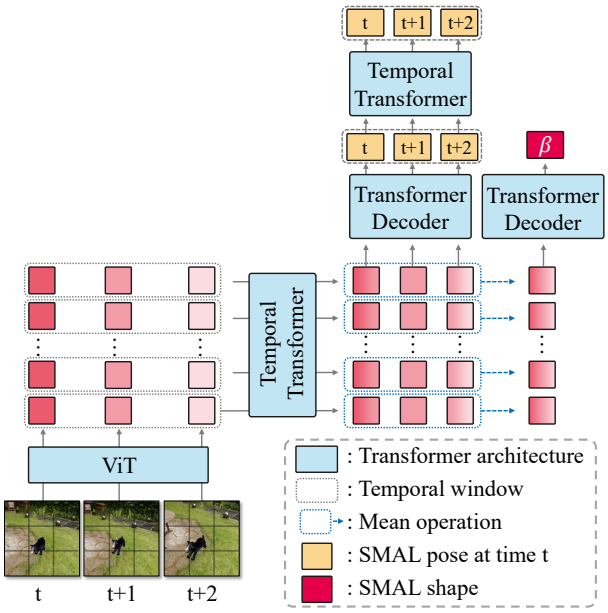

Figure 3. **Animal Video Transformer (AVT) architecture.** Temporal modeling is performed after the ViT backbone. The transformer decoder predicts frame-wise SMAL poses and translations together with a sequence-level shape parameter.

parameters $\{\hat{\theta}_t\}_{t=1}^{T}$ and root translations $\{\hat{T}_t\}_{t=1}^{T}$ are estimated for each frame. This design explicitly separates identity from motion and suppresses frame-wise shape drift.

To process long videos, AVT operates on sliding temporal windows of length 16 with stride 1. Each window is processed independently, and predictions from overlapping windows are aggregated to obtain the final per-frame outputs. This design allows AVT to model local temporal context while remaining applicable to arbitrary input videos.

We freeze the pre-trained ViT backbone to preserve strong visual features and train only the temporal transformer and decoder on our synthetic video dataset. The model is optimized with the following objective:

$$\mathcal{L} = \lambda_{2D}\mathcal{L}_{2D} + \lambda_{3D}\mathcal{L}_{3D} + \lambda_{SMAL}\mathcal{L}_{SMAL} + \lambda_{V}\mathcal{L}_{V}, \quad (3)$$

where $\mathcal{L}_{2D}$ is the 2D keypoint reprojection loss, $\mathcal{L}_{3D}$ is the 3D keypoint loss, $\mathcal{L}_{SMAL}$ is the loss on the regressed SMAL parameters $(\hat{\beta}, \hat{\theta})$, and $\mathcal{L}_{V}$ is the vertex loss on the final mesh. We set $\lambda_{2D} = \lambda_{3D} = 5$ and $\lambda_{SMAL} = \lambda_{V} = 1$.

4. Experiments

We evaluate AVT on both synthetic and in-the-wild videos. We first describe the experimental setup, then compare against strong per-frame baselines, and finally analyze each component through ablations.

4.1. Experimental Setup

Datasets. We primarily evaluate on WildAni4D, our synthetic dataset of 30K video sequences spanning 56 animal species, with ground-truth SMAL parameters, 3D joint locations, and camera motion for every frame. As detailed in Sec. 3.2, this dataset enables quantitative evaluation of temporal 4D animal reconstruction. Detailed train/test splits are provided in the supplementary material. For real-world evaluation, where 3D ground truth is unavailable, we use a curated set of in-the-wild videos with challenging animal motion (Cat, Lion, and Rhino).

Evaluation Metrics. We report standard pose metrics, including MPJPE, S-MPJPE, and PA-MPJPE, together with two temporal metrics. The first is acceleration error (Accel), computed as the L2 difference between the ground-truth and predicted joint accelerations, averaged over joints and frames. The second is shape consistency (SC), which measures frame-to-frame variation in the predicted SMAL shape parameters:

$$\text{SC} = \frac{1}{T-1}\sum_{t=1}^{T-1} \left\| \boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1} \right\|_2 . \quad (4)$$

Lower SC indicates more stable and physically plausible identity estimates over time.

Although AVT predicts a sequence-level shape parameter within each temporal window, the shape consistency metric can still be non-zero in practice. This is because we process the long video using overlapping sliding windows, and aggregating predictions from windows can introduce small frame-wise variations in the final β_t estimates.

For in-the-wild evaluation, we adopt a reconstruction-based protocol. We use GART [16] to reconstruct an animatable 3D avatar from the input video and drive it with the pose sequence predicted by each method. We then compare the rendered output against the input frames using PSNR, SSIM, and LPIPS [45]. Higher PSNR/SSIM and lower LPIPS indicate better reconstruction fidelity.

Baselines. We compare against two strong ViT-based per-frame animal mesh recovery methods, AniMer [21] and GenZoo [23]. We evaluate each method in two settings. AniMer* and GenZoo* denote the official pretrained models evaluated directly on our test set. AniMer and GenZoo denote baselines obtained by fine-tuning the models on WildAni4D for 10,000 iterations with batch size 16, which takes about 24 hours on a single NVIDIA 3090 Ti GPU.

For the in-the-wild reconstruction benchmark, we compare Ours-tto + GART against AniMer* + GART and GenZoo* + GART. Here, Ours-tto denotes test-time optimization over pose and translation, without shape.

Implementation Details. AVT uses the HMR2.0 ViT backbone pretrained on GenZoo [23]. Following

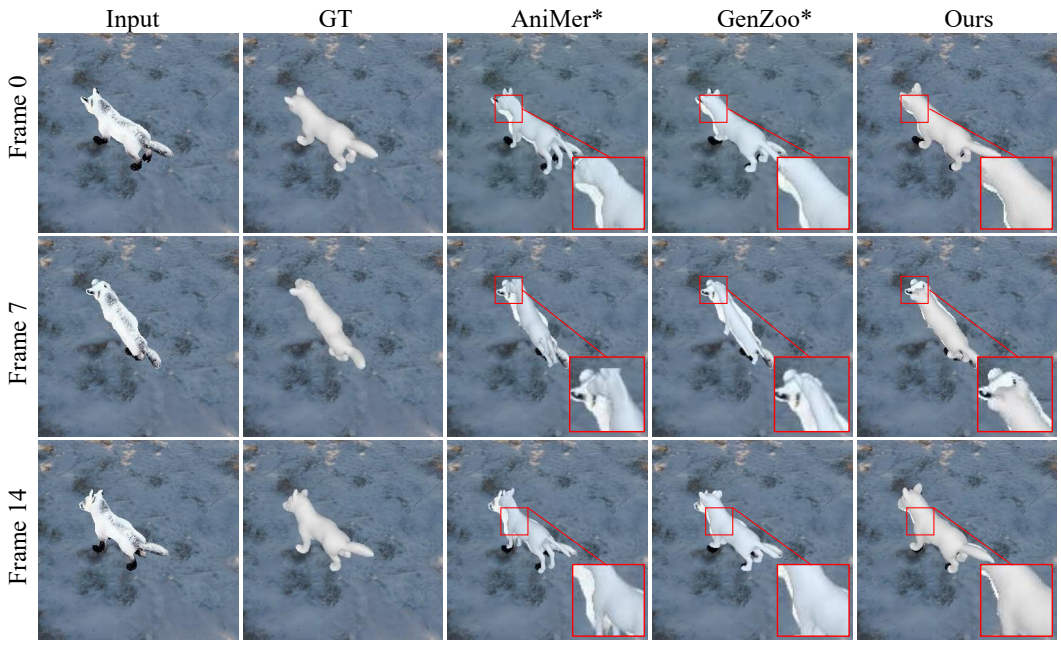

Figure 4. **Qualitative comparison on the WildAni4D test set.** We compare reconstructions across multiple frames from the same video. The highlighted regions show that per-frame baselines exhibit pose-dependent shape drift and temporal inconsistency, whereas our method preserves more stable body proportions and local contours over time.

Table 1. **Quantitative comparison on the WildAni4D test set.** Asterisks ($*$) indicate official pretrained models. We compare pose accuracy and temporal consistency against both official pretrained baselines and their fine-tuned variants.

	MPJPE↓	S-MPJPE↓	PA-MPJPE↓	Accel↓	Shape Consistency↓
AniMer*	-	-	79.02	55.92	0.4027
GenZoo*	-	-	82.78	58.75	0.2307
AniMer	123.42	81.79	53.91	19.01	0.4215
GenZoo	109.23	69.66	40.35	19.37	0.4487
Ours	**91.29**	**62.44**	**40.13**	**7.35**	**0.0703**

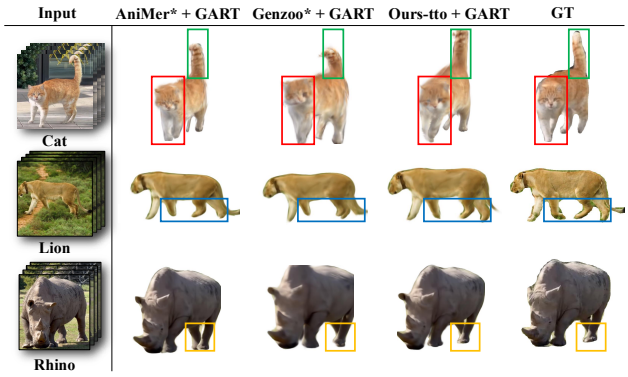

Figure 5. **Qualitative results on the in-the-wild dataset.** We optimize GART using the pose predictions of each method and render the resulting avatar to the target frame. The colored boxes highlight that our method better matches the ground-truth pose, particularly in local body orientation and limb configuration.

TRAM [35], we freeze the ViT backbone and train the temporal transformer and decoder from scratch. We train for 60,000 iterations with a sequence length of 16 and a batch size of 12. Training takes approximately 24 hours on a single NVIDIA 3090 Ti GPU. We use AdamW [15, 20] with learning rates of 1×10^{-5} for the decoder and 3×10^{-5} for the temporal transformer.

4.2. Comparison to Animal Mesh Recovery

Qualitative Results. Figure 4 reveals two characteristic failure modes of per-frame baselines: temporal shape drift and pose-dependent identity changes. Although all methods process the same animal sequence, AniMer* and GenZoo* produce visibly inconsistent body proportions across Frames 0, 7, and 14. This is particularly evident in the zoomed regions, where the local silhouette and body thickness vary from frame to frame, even though the underlying animal identity should remain fixed throughout the video. As a result, the reconstructed animal appears to change its shape as the pose changes, indicating insufficient disentanglement between pose and shape. In contrast, our method preserves a much more stable body shape over time, maintaining consistent overall proportions and local contours while still capturing large articulated motion. This visual stability suggests that our sequence-level shape mod-

Table 2. **Quantitative results on the in-the-wild dataset.** We report photometric and perceptual metrics on rendered target views.

Data	Method	PSNR ↑	SSIM ↑	LPIPS ↓
Cat	AniMer* + GART	18.39	0.8186	0.2243
	GenZoo* + GART	18.03	0.8154	0.2568
	Ours + GART	17.97	0.8203	0.2460
	Ours-tto + GART	**18.60**	**0.8256**	**0.2231**
Lion	AniMer* + GART	15.52	0.8516	0.2657
	GenZoo* + GART	16.80	**0.8680**	0.2134
	Ours + GART	16.40	0.8664	0.2314
	Ours-tto + GART	**17.83**	0.8395	**0.2049**
Rhino	AniMer* + GART	14.99	0.7426	0.2614
	GenZoo* + GART	11.05	0.6576	0.4499
	Ours + GART	12.94	0.7014	0.3427
	Ours-tto + GART	**16.31**	**0.7678**	**0.2439**

eling effectively suppresses frame-wise identity drift and yields temporally coherent 4D reconstruction.

Quantitative Results. Table 1 shows that our method achieves the best overall performance. Relative to the strongest fine-tuned baseline, it reduces Shape Consistency from 0.4215 to 0.0703 and Accel from 19.01 to 7.35, while also improving S-MPJPE and PA-MPJPE. These results indicate that enforcing sequence-level shape consistency substantially improves temporal stability.

4.3. Comparison to Animatable Reconstruction

Although AVT enables controlled evaluation of temporal consistency, real-world validation remains important. Since 3D ground truth is unavailable for in-the-wild videos, we evaluate the utility of the predicted pose sequences in a downstream animatable reconstruction setting using GART.

Qualitative Results. Figure 5 shows that the quality of animatable reconstruction strongly depends on how accurately the predicted pose drives GART. Since GART renders the reconstructed avatar using the estimated pose sequence, better pose estimation should produce renderings that more faithfully match the target frame. In the examples shown, AniMer* + GART and GenZoo* + GART exhibit noticeable pose misalignment, highlighted by the red and yellow boxes, including inaccurate torso orientation and limb placement for the Cat, leg configuration errors for the Lion, and mismatched local pose details for the Rhino. In contrast, Ours-tto + GART more faithfully reproduces the target pose and yields renderings that are visually closer to the ground-truth frame. These results suggest that the improved pose accuracy of our method provides a stronger motion signal for downstream animatable reconstruction.

Quantitative Results. Table 2 shows that Ours-tto + GART achieves the strongest overall performance across the three sequences, especially in PSNR and LPIPS, while remaining competitive on SSIM. These trends are consistent across Cat, Lion, and Rhino, indicating that the improved temporal stability of our pose estimates generally benefits downstream animatable reconstruction in in-the-wild videos.

Table 3. **Ablation study.** Contribution of each architectural component. We analyze the effect of temporal modeling and sequence-level shape regression on pose accuracy and temporal stability.

	MPJPE↓	S-MPJPE↓	PA-MPJPE↓	Accel↓	Shape Consistency↓
AVT-a	109.23	69.66	40.35	19.37	0.4487
AVT-b	95.71	**62.36**	40.51	19.21	0.5657
AVT (Ours)	**91.29**	62.44	**40.13**	**7.35**	**0.0703**

4.4. Ablation Studies on Animal Mesh Recovery

We analyze the contribution of each component in Table 3. AVT-a is the HMR2.0 ViT architecture retrained on WildAni4D. It already provides reasonable pose accuracy but suffers from large temporal error and severe shape instability. AVT-b adds a temporal transformer, which improves pose estimation but further worsens shape consistency. This shows that temporal modeling alone does not solve the identity-drift problem. In contrast, adding the sequence-level shape regressor reduces Shape Consistency from 0.5657 to 0.0703 and Accel from 19.21 to 7.35, while also improving MPJPE and PA-MPJPE. These results show that the sequence-level shape constraint is the key component for stable 4D animal reconstruction. Additional ablations are provided in the supplementary material.

4.5. Applications

Our method enables several practical downstream applications built on temporally consistent 4D animal reconstruction, as illustrated in detail in Fig. 6. In particular, the recovered motion can be used for pseudo-GT annotation, animatable animal reconstruction, and text-to-motion generation.

Pseudo-GT Annotation. AVT provides stable initialization for unlabeled in-the-wild videos. By fixing the sequence-level shape and optimizing only pose and translation with 2D keypoints, we obtain temporally coherent pseudo-ground-truth motion suitable for motion annotation.

Animatable Animal Reconstruction. Our predictions can be used to optimize explicit animatable models such as GART, yielding controllable 3D animal avatars. This provides a practical interface between monocular 4D animal reconstruction and downstream animation pipelines: a video can be converted into a pose-driven 3D animal representation that can be replayed, edited, and rendered from new viewpoints. Such a representation may serve as a useful starting point for future animatable animal modeling and content creation systems.

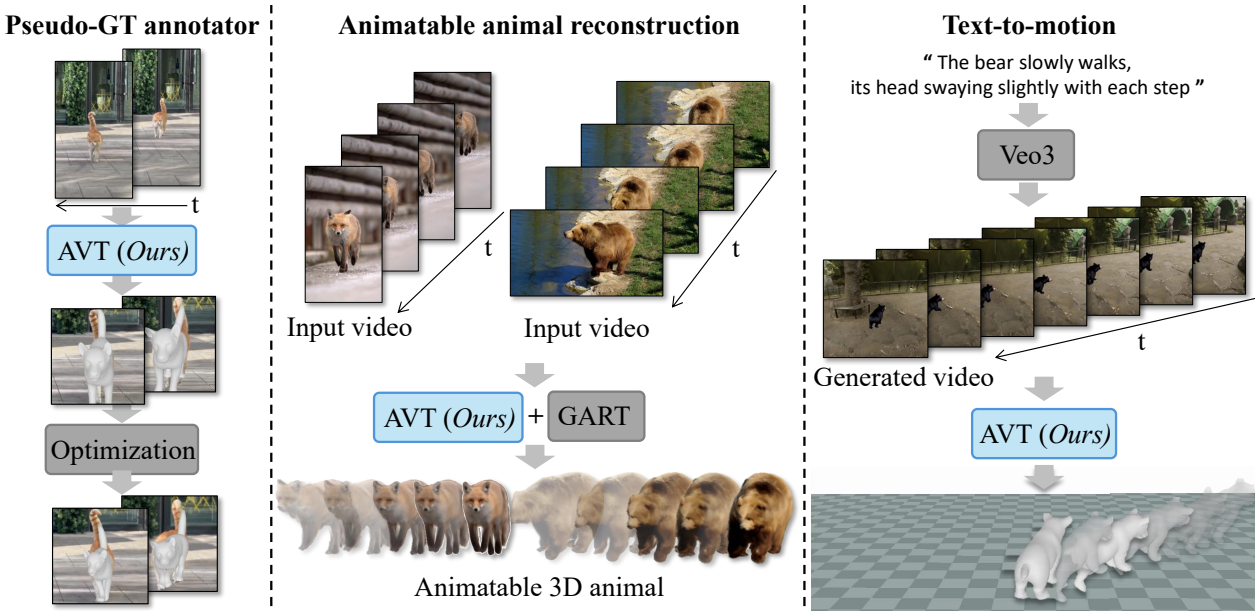

Figure 6. **Applications of WildAni4D.** Our reconstructed 4D animal motions enable multiple downstream applications, including pseudo-GT annotation, animatable animal reconstruction, and text-to-motion generation. **Left:** AVT provides temporally coherent initialization that can be further refined for pseudo-GT annotation. **Middle:** AVT predictions can be used to optimize GART and obtain controllable, animatable 3D animal avatars. **Right:** AVT can lift text-to-video generations into plausible 3D animal motion sequences.

Text-to-Motion. AVT can also lift generated animal videos into plausible 3D motion sequences. In particular, by generating a video from a text prompt using a text-to-video model such as GOOGLE VEO3 [9] and then applying AVT, we obtain a text-conditioned 3D motion sequence without requiring manual 3D annotation. This suggests a promising direction for building text-motion paired animal datasets at scale, which could support future research on animal motion generation, retrieval, and text-conditioned animation.

5. Conclusion and Discussion

In this work, we propose WildAni4D, a unified framework for 4D animal mesh reconstruction from monocular in-the-wild videos. WildAni4D combines a scalable synthetic video generation pipeline with a temporally aware reconstruction model for temporally coherent and metric-scale recovery of animal motion. The synthetic pipeline generates realistic annotated training videos with dynamic textured animals, diverse 3D scenes, and simulated camera trajectories. The reconstruction model disentangles camera motion from animal motion and enforces sequence-level shape consistency for stable 4D reconstruction over time. Experiments on synthetic and in-the-wild videos show improved pose accuracy and temporal stability over per-frame baselines. WildAni4D also supports downstream applications, including pseudo-ground-truth annotation, animatable animal reconstruction, and text-to-motion generation.

Limitations. Our framework relies heavily on synthetic supervision for 4D animal motion reconstruction. Synthetic data cannot fully reflect the visual complexity, motion diversity, and capture conditions of real-world animal videos. As a result, the model may struggle to generalize under challenging in-the-wild scenarios that deviate from the training distribution. We plan to reduce this gap by developing more realistic and diverse training data and by improving generalization to real-world videos.

Acknowledgments. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2020-II201336, Artificial Intelligence Graduate School Program (UNIST); No.RS-2022-II220612, Geometric and Physical Commonsense Reasoning based Behavior Intelligence for Embodied AI; No.RS-2025-25442149, LG AI STAR Talent Development Program for Leading Large-Scale Generative AI Models in the Physical AI Domain), by the InnoCORE program of the Ministry of Science and ICT (25-InnoCORE-01), by computing support on the Vista GPU Cluster through the Center for Generative AI (CGAI) and the Texas Advanced Computing Center (TACC) at the University of Texas at Austin.

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 4

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 4

[3] 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*, 2020. 2

[4] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, 2023. 2

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, 2017. 2, 4

[6] Gyeongsu Cho, Changwoo Kang, Donghyeon Soon, and Kyungdon Joo. Dogrecon: Canine prior-guided animatable 3d gaussian dog reconstruction from a single image: G. cho et al. *IJCV*, 133(9):6332–6346, 2025. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[8] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2, 3, 4

[9] Google DeepMind. Veo. https://deepmind.google/models/veo/. 8

[10] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *CVPR*, 2024. 2, 4

[11] 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, 2013. 2

[12] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3

[13] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *ICCV*, 2023. 3

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 2, 4

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[16] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *CVPR*, 2024. 2, 5

[17] Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 4dcomplete: Non-rigid motion estimation beyond the observable surface. In *ICCV*, 2021. 2

[18] 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*, 2024. 2

[19] Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion. In *SIGGRAPH Asia*, 2024. 4

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[21] Jin Lyu, Tianyi Zhu, Yi Gu, Li Lin, Pujin Cheng, Yebin Liu, Xiaoying Tang, and Liang An. Animer: Animal pose and shape estimation using family aware transformer. In *CVPR*, 2025. 2, 3, 4, 5

[22] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 2, 3

[23] Tomasz Niewiadomski, Anastasios Yiannakidis, Hanz Cuevas-Velasquez, Soubhik Sanyal, Michael J Black, Silvia Zuffi, and Peter Kulits. Generative zoo. In *ICCV*, 2025. 2, 3, 4, 5

[24] 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*, 2021. 2

[25] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 4

[26] Nadine Rueegg, Silvia Zuffi, Konrad Schindler, and Michael J Black. Barc: Learning to regress 3d dog shape from images by exploiting breed information. In *CVPR*, 2022. 2

[27] Nadine Rüegg, Shashank Tripathi, Konrad Schindler, Michael J Black, and Silvia Zuffi. Bite: Beyond priors for improved three-d dog pose estimation. In *CVPR*, 2023. 2

[28] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia*, 2024. 2, 3

[29] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *CVPR*, 2024. 3

[30] Moira Shooter, Charles Malleson, and Adrian Hilton. Sydog: A synthetic dog dataset for improved 2d pose estimation. *arXiv preprint arXiv:2108.00249*, 2021. 3

[31] Moira Shooter, Charles Malleson, and Adrian Hilton. Digidogs: Single-view 3d pose estimation of dogs using synthetic training data. In *WACV*, 2024. 3

[32] 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

[33] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In *NeurIPS*, 2021. 3, 4

[34] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 3

[35] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *ECCV*, 2024. 2, 3, 4, 6

[36] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *CVPR*, 2023. 2

[37] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 2, 4

[38] 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*, 2022. 2

[39] Gengshan Yang, Chaoyang Wang, N Dinesh Reddy, and Deva Ramanan. Reconstructing animatable categories from videos. In *CVPR*, 2023. 2

[40] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, et al. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *ICCV*, 2023. 2

[41] Zhangsihao Yang, Mingyuan Zhou, Mengyi Shan, Bingbing Wen, Ziwei Xuan, Mitch Hill, Junjie Bai, Guo-Jun Qi, and Yalin Wang. Omnimotiongpt: Animal motion generation with limited data. In *CVPR*, 2024. 2, 3

[42] 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*, 2022. 2

[43] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023. 3

[44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[46] Silvia Zuffi and Michael J Black. Awol: Analysis without synthesis using language. In *ECCV*, 2024. 2, 3

[47] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *CVPR*, 2017. 2, 3

[48] 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*, 2024. 2