

AniGauss: Toward Animatable Animal Reconstruction from Single In-the-Wild Images via Topology-Aware Gaussians

Gyeongsu Cho¹ Changwoo Kang¹ Donghyeon Soon³ Hezhen Hu² Kyungdon Joo^{1*}

¹UNIST ²University of Texas at Austin ³DGIST

{threedv, kangchangwoo, kyungdon}@unist.ac.kr

dhssoon@dgist.ac.kr alexhu@utexas.edu

Abstract

*We address feed-forward animatable animal reconstruction from a single in-the-wild image, a challenging setting where the input is severely under-constrained yet the output must remain both visually faithful and controllable. Our framework decomposes the task into two stages: a canonical four-view generation module that predicts front, left, right, and back views of the same animal, and a topology-aware reconstruction module that converts these views into a SMAL-aligned animatable Gaussian representation. By fusing multi-view appearance cues in a UV-aligned canonical space and anchoring Gaussian primitives to the articulated animal topology, our method produces reconstructions that preserve both structural coherence and re-poseability. To support this setting, we build synthetic supervision from textured SMAL animal assets with exact canonical-view and articulated 3D annotations. Experiments on challenging in-the-wild quadruped images suggest that our approach produces more consistent novel-view renderings and more plausible animatable reconstructions than generic image-to-3D and animal reconstruction baselines. **Project page:** <https://vision3d-lab.github.io/anigauss/>.*

1. Introduction

Reconstructing a controllable 3D animal from visual observations is a long-standing goal in computer vision and graphics. Such a representation would support applications including pose-driven animation, content creation, virtual production, and animal behavior analysis. Although recent advances in generative 3D modeling and neural rendering have improved single-image 3D reconstruction, most methods still recover only static geometry or category-agnostic appearance, without a structured mechanism for animation and control. Yet many downstream applications require not

only reconstructing an animal’s appearance, but also recovering a representation that can be reposed and rendered consistently under novel articulations [19, 29, 30].

This challenge is especially severe for animals captured in the wild. Compared with humans [14, 32], animals show much larger shape variation across species and breeds, more diverse fur patterns and textures, and frequent self-occlusion under non-cooperative poses and viewpoints. In-the-wild animal images are also captured from arbitrary viewpoints under uncontrolled lighting and cluttered backgrounds. These factors make single-image animatable reconstruction fundamentally difficult: the problem is highly under-constrained, yet the output must preserve both appearance fidelity and articulation-aware structure. Existing methods do not fully address this setting [11, 12, 16–18]. Some recover animatable representations from video using temporal consistency and motion cues, but they rely on multi-frame input and do not apply to single-image inference. Others operate on single images, but are either limited to narrow categories such as dogs or recover only coarse pose and shape without a textured, animatable output.

In this paper, we study the underexplored setting of feed-forward animatable animal reconstruction from a single in-the-wild image. Our goal is to reconstruct a 3D animal representation that is both visually faithful and controllable under a parametric animal prior [33, 34], enabling downstream animation from a single casually captured photograph. To make this feasible, we decompose the problem into two stages. First, from one input image, we infer a small set of semantically consistent canonical views that expose the animal from multiple directions. Second, we reconstruct an animatable 3D representation from these views using a topology-aware Gaussian formulation aligned with the SMAL prior. This design lets the model exploit strong multi-view reconstruction machinery and produce an output compatible with articulated control.

Our key idea is to build animatability into the representation itself. Instead of predicting unconstrained 3D primitives or a static radiance field, we anchor Gaussian primi-

*Corresponding author.

tives to a deformable animal topology and decode them in a canonical space aligned with the body prior [5]. Concretely, our framework combines a multi-view image encoder with a SMAL-aware Gaussian decoder that predicts appearance-aware Gaussian attributes while preserving explicit correspondence to an articulated quadruped template. As a result, the reconstructed animal can be reposed through the underlying parametric model while maintaining coherent appearance and geometry. This representation bridges recent feed-forward 3D reconstruction pipelines, which are efficient but usually static, and model-based animal capture methods, which are controllable but typically require video or optimization-heavy inference.

Our formulation differs from prior work in several ways. Compared with video-based animal reconstruction methods [19, 29, 30], we remove the dependence on motion and temporal observations and operate in the harder single-image regime. Compared with category-specific single-image approaches, we target a broader family of quadruped animals rather than one species. Compared with methods that estimate only pose or coarse shape, we recover an animatable appearance representation for controllable rendering. More broadly, our method suggests that single-image animal reconstruction need not be limited to static 3D recovery; by coupling feed-forward view completion with topology-aware articulated decoding, we can reconstruct a controllable animal avatar from one image.

To support this setting, we also build on AniGA10k, a dataset for animatable animal reconstruction designed to provide multi-view and articulation-aware supervision for quadruped animals. It supports controllable reconstruction beyond narrow species categories.

We evaluate our approach against strong baselines for image-to-3D and animal reconstruction, including generic feed-forward 3D methods and animal-adapted animatable reconstruction baselines. Experiments show that our method produces higher-quality controllable reconstructions, with better appearance fidelity and more reliable pose control in challenging in-the-wild cases.

In summary, our contributions are threefold.

- We formulate single-image animatable animal reconstruction in the wild as a two-stage problem that separates canonical view completion from animatable 3D recovery.
- We introduce a topology-aware Gaussian representation anchored to the SMAL prior, enabling pose-conditioned articulation after reconstruction.
- We build a synthetic supervision pipeline and provide qualitative evidence on in-the-wild quadruped images.

2. Related Work

Animatable animal and human avatars. Animatable reconstruction methods recover controllable dynamic objects using motion, temporal consistency, or body priors. For

animals, video-based methods recover deformable models from casual sequences [10, 19, 21, 28–30], but they rely on multi-frame evidence and often require per-sequence optimization. For humans, neural and Gaussian avatar methods have shown that body-model priors can support pose-controlled rendering [3, 4, 6, 7, 13, 14, 25, 32]. However, these approaches are typically designed for human topology, calibrated or video observations, or person-specific capture. We target the complementary setting of feed-forward animatable reconstruction for quadruped animals from a single in-the-wild image.

Single-image 3D generation and view completion. Recent image-to-3D methods have improved single-view reconstruction by combining large reconstruction networks, diffusion-based view synthesis, and efficient 3D representations [1, 8, 20, 23, 24, 27, 31]. Gaussian splatting further enables efficient radiance-field-like rendering with explicit primitives [5, 15, 22]. Despite their visual quality, most generic image-to-3D methods produce static assets and do not expose an articulation model for reliable pose control. Our pipeline separates semantic view canonicalization from topology-aware Gaussian decoding, keeping the output compatible with SMAL-driven articulation.

3. Method

Our goal is to reconstruct an *animatable* 3D animal avatar from a single in-the-wild image in a feed-forward manner. This problem is challenging because a single image provides only partial appearance and geometry, whereas controllable reconstruction requires both multi-view-consistent shape and explicit articulation-aware structure. We address this challenge with a two-stage formulation. First, we transform the input image into a compact set of canonical semantic views. Second, we reconstruct an animatable animal representation from these views using a topology-aware Gaussian model aligned with the SMAL prior.

Formally, given an input image I^{wild} , we first predict four canonical views

$$\hat{\mathcal{V}} = \left\{ \hat{I}^{\text{front}}, \hat{I}^{\text{left}}, \hat{I}^{\text{right}}, \hat{I}^{\text{back}} \right\}, \quad (1)$$

and then reconstruct an animatable animal representation $\mathcal{A}$:

$$\mathcal{A} = \mathcal{F}_{\theta}(\hat{\mathcal{V}}). \quad (2)$$

The first stage reduces viewpoint ambiguity by canonicalizing the observation, while the second stage ensures that the recovered representation is not merely static, but explicitly controllable under articulated motion.

3.1. Synthetic Training Data Construction

A core obstacle in this task is the lack of supervision that simultaneously provides (i) an in-the-wild-style monocular

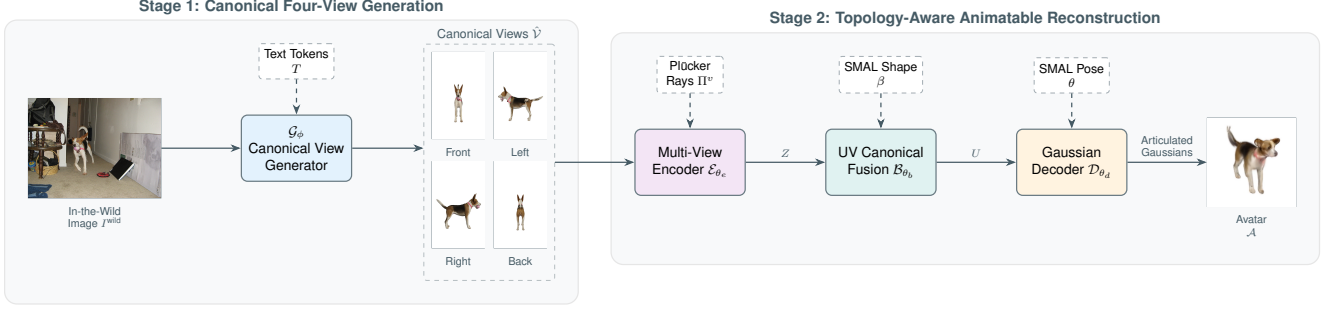


Figure 1. Overview of our two-stage framework. Given a single in-the-wild image I^{wild} , we first generate four canonical semantic views $\hat{\mathcal{V}} = \{\hat{I}^{\text{front}}, \hat{I}^{\text{left}}, \hat{I}^{\text{right}}, \hat{I}^{\text{back}}\}$ using a LoRA-adapted canonical view generator. We then reconstruct an animatable animal avatar $\mathcal{A}$ by encoding the ordered four-view observations, fusing them in a UV-aligned canonical space, and decoding SMAL-topology-anchored articulated Gaussians.

input, (ii) canonical multi-view supervision, and (iii) exact articulated 3D ground truth. To address this, we build synthetic data from controlled textured SMAL+ [33] meshes.

We generate textured animal identities by combining a parametric animal body prior with text-guided shape and texture synthesis, using AWOL [33] to obtain expressive SMAL+ animal instances and SyncMVD [9] to generate appearance-consistent textured meshes. Because shape, articulation, and texture are all known, the resulting assets provide precise 3D supervision during rendering. To avoid training only on canonical poses, we extract global orientation and body pose statistics from real animal videos and use them to randomly articulate the generated animals before rendering the input image. This produces monocular inputs resembling arbitrary in-the-wild observations while preserving exact ground-truth geometry and pose. For each posed input image, we render the same animal identity in a canonical reference pose and camera setup, sample 24 views over 360° , and select front, left, right, and back as canonical targets with exact SMAL+ supervision.

Overall, our training pairs take the form

$$I^{\text{wild}} \rightarrow \{I^{\text{front}}, I^{\text{left}}, I^{\text{right}}, I^{\text{back}}\} \rightarrow \mathcal{A}^*, \quad (3)$$

where I^{wild} is an in-the-wild-style monocular input, the canonical images correspond to the same animal identity rendered under controlled viewpoints, and $\mathcal{A}^*$ denotes the target animatable 3D supervision.

3.2. Canonical Four-View Generation

Given an arbitrary in-the-wild animal image, our first stage predicts four canonical images corresponding to the front, left, right, and back views of the same identity. We implement this stage by fine-tuning an image editing model via LoRA adaptation [2, 26].

The objective of this stage is not generic editing, but *semantic viewpoint canonicalization*. During training, the model receives an in-the-wild-style image together

with a learned view token, e.g., `<animal4v> frontal view`, and is optimized to produce the corresponding canonical target image:

$$\hat{I}^v = \mathcal{G}_\phi(I^{\text{wild}}, t^v), \quad v \in \{\text{front, left, right, back}\}, \quad (4)$$

where $\mathcal{G}_\phi$ denotes the LoRA-adapted image editing model and t^v is the learned view-specific token.

This design has two advantages. First, it converts a highly unconstrained monocular observation into an ordered set of reconstruction-friendly canonical views. Second, because the model is trained on exact synthetic input-target pairs generated from textured SMAL+ animals, it learns animal-specific structural regularities rather than generic object-level view completion.

3.3. Topology-Aware Four-View Animatable Reconstruction

Given the four canonical views, our second stage reconstructs an animatable animal representation. The reconstruction network is designed around three principles: (i) semantic multi-view encoding, (ii) canonical-space fusion, and (iii) template-aware animatable decoding.

3.3.1. Geometry-Aware Multi-View Encoding

We encode the ordered four-view inputs with a geometry-aware multi-view encoder. For each view, we concatenate the RGB image with Plücker ray embeddings derived from the corresponding camera parameters. Let

$$X^v = [I^v, \Pi^v], \quad (5)$$

where Π^v denotes the ray embedding for view v . The four-view tensor is then processed by a shared multi-view encoder:

$$Z = \mathcal{E}_{\theta_e}(X^{\text{front}}, X^{\text{left}}, X^{\text{right}}, X^{\text{back}}). \quad (6)$$

This representation is semantically structured rather than unordered. Because the four inputs always correspond

to fixed canonical directions, the encoder can learn stable cross-view correspondences across the canonical views.

3.3.2. UV-Space Canonical Fusion

Instead of decoding directly into unconstrained 3D primitives, we first lift the encoded multi-view representation into a canonical UV-aligned latent space:

$$U = \mathcal{B}_{\theta_b}(Z, \beta), \quad (7)$$

where $\mathcal{B}_{\theta_b}$ denotes the UV fusion module and β denotes the SMAL shape parameters. The resulting UV latent acts as a structured atlas on the canonical animal surface.

The UV-space design is important because it allows appearance information from multiple views to be accumulated in a topology-aware coordinate system. As a result, the decoder does not need to infer correspondence from scratch; instead, it receives a latent representation already aligned with the deformable animal body.

3.3.3. SMAL-Topology-Anchored Gaussian Decoding

From the UV latent U , we decode a set of Gaussian primitives anchored to the SMAL topology. Let $\mathbf{v}_i$ denote the i -th anchor point sampled on the canonical SMAL surface. For each anchor, the decoder predicts a local offset $\Delta\mathbf{v}_i$, opacity α_i , color $\mathbf{c}_i$, scale $\mathbf{s}_i$, and local rotation $\mathbf{R}_i^{\text{loc}}$:

$$(\Delta\mathbf{v}_i, \alpha_i, \mathbf{c}_i, \mathbf{s}_i, \mathbf{R}_i^{\text{loc}}) = \mathcal{D}_{\theta_d}(U, i). \quad (8)$$

The canonical Gaussian center is defined as

$$\hat{\mathbf{v}}_i = \mathbf{v}_i + \Delta\mathbf{v}_i. \quad (9)$$

Given SMAL pose parameters θ and shape parameters β , each anchor is articulated by the corresponding skinning transform $T_i(\theta, \beta)$. The final Gaussian center is

$$\mathbf{x}_i = T_i(\theta, \beta) \hat{\mathbf{v}}_i. \quad (10)$$

Likewise, the final Gaussian orientation is obtained by composing the articulated template rotation with the predicted local rotation:

$$\mathbf{R}_i = \mathbf{R}_i^{\text{art}}(\theta, \beta) \mathbf{R}_i^{\text{loc}}. \quad (11)$$

The resulting Gaussian state is therefore not a free-floating point cloud, but a template-aligned structured representation. Because the Gaussian anchors are explicitly tied to the SMAL body, the reconstructed animal remains animatable under novel poses after inference.

3.4. Rendering and Training Objectives

We train the reconstruction network with rendering-based supervision. Given the predicted Gaussian representation and the target SMAL parameters, we render the animal

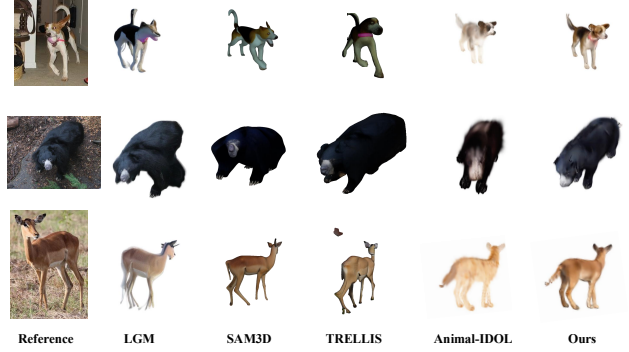


Figure 2. Qualitative single-image-to-3D comparison.

from the supervised camera views and compare the rendered outputs with the ground-truth canonical target images. The training objective combines pixel-level reconstruction losses, an optional perceptual loss, a head-focused loss for preserving identity-sensitive facial details, silhouette supervision when masks are available, and regularization terms on Gaussian attributes such as opacity, scale, and deformation offsets. Together, these objectives encourage globally coherent reconstruction while adding extra supervision for local regions such as the head, legs, and tail.

4. Experiments

We evaluate our method through qualitative comparisons with representative image-to-3D baselines. Since there is no established benchmark for single-image animatable animal reconstruction in the wild, this paper focuses on visual plausibility under a shared single-image input setting on challenging quadruped images.

4.1. Baselines

We use strong generic image-to-3D baselines (LGM [23], SAM3D [24], TRELLIS [27]). We also include an animal-adapted version of IDOL [32] as an animatable baseline. Since the original IDOL is designed for human reconstruction, we retrain it on our AniGA10k-based animal data to obtain a quadruped baseline with identical inputs.

4.2. Qualitative Comparison on Image to 3D

Figure 2 compares our method against LGM, SAM3D, TRELLIS, and the animal version of IDOL on challenging in-the-wild quadruped images. These examples include significant articulation variation, partial self-occlusion, fur texture, and complex silhouettes.

Compared with generic image-to-3D baselines, our method produces reconstructions with stronger animal-specific structural consistency and more faithful appearance preservation. In particular, LGM, SAM3D, and TRELLIS often generate plausible coarse geometry, but they tend to struggle with consistent limbs, plausible rear views, and

identity-preserving texture completion under large view-point changes. Their outputs may look visually reasonable from limited viewpoints, yet often become inconsistent when rendered from unseen views, especially for thin legs, tails, and head geometry.

Compared with Animal-IDOL, our method produces more stable multi-view appearance and sharper animal-specific structural details. This improvement is particularly visible in regions that are difficult to infer from a single image, such as the back, side silhouette, and partially occluded limbs. We hypothesize that this behavior is related to two design choices: (i) the explicit canonical-view generation stage, which may reduce monocular ambiguity, and (ii) the topology-aware reconstruction design, which ties the recovered representation to the articulated animal prior.

Overall, the qualitative comparisons suggest that our method produces more consistent novel-view renderings and more plausible full-body reconstructions than these baselines in this challenging setting.

5. Conclusion

We introduced AniGauss, a feed-forward framework for reconstructing animatable animal avatars from a single in-the-wild image. Our approach first predicts canonical semantic views and then converts them into a SMAL-topology-aware Gaussian representation, enabling visually consistent novel-view rendering and pose-conditioned animation. The qualitative comparisons suggest that canonical-view generation and topology-aware reconstruction provide a practical path toward single-image animatable animal reconstruction.

Limitations. Our current validation is primarily qualitative, since there is no established real-image benchmark for single-image animatable animal reconstruction in the wild. In addition, our method is built on the SMAL [34] topology and therefore works best for quadruped animals that can be reasonably aligned to this prior. Animals with geometry that significantly deviates from SMAL, such as prominent horns or category-specific structures outside the template space, remain challenging for the current framework.

Future Work. An important direction for future work is to build a real-image benchmark for animatable animal reconstruction, with diverse species, poses, viewpoints, and appearance variations, together with standardized evaluation protocols. Another promising direction is to extend the current SMAL-based formulation toward broader animal priors and more flexible representations, so that the method can better handle animals beyond the current quadruped template family. We also plan to analyze the contributions of canonical-view generation and topology-aware reconstruction, and to study failures under heavy occlusion,

unusual viewpoints, and out-of-distribution appearances in the wild.

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).

References

- [1] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2
- [2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022. 3
- [3] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. *arXiv preprint arXiv:2312.02134*, 2023. 2
- [4] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, and Deng Cai. One-shot implicit animatable avatars with model-based priors. In *ICCV*, 2023. 2
- [5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. 2
- [6] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. *arXiv preprint arXiv:2311.17910*, 2023. 2
- [7] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *CVPR*, 2024. 2
- [8] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2
- [9] Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. 3
- [10] Haimin Luo, Teng Xu, Yuheng Jiang, Chenglin Zhou, Qiwei Qiu, Yingliang Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Artemis: Articulated neural pets with appearance and motion synthesis. *ACM Transactions on Graphics*, 41(4), 2022. 2
- [11] 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*, pages 17486–17496, 2025. 1
- [12] Tomasz Niewiadomski, Anastasios Yiannakidis, Hanz Cuevas-Velasquez, Soubhik Sanyal, Michael J Black, Silvia Zuffi, and Peter Kulits. Generative zoo. In *ICCV*, pages 8492–8502, 2025. 1
- [13] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 2
- [14] Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu, Zhe Li, Weihao Yuan, Liefeng Bo, et al. Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21148–21158, 2025. 1, 2
- [15] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 2
- [16] 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*, pages 3876–3884, 2022. 1
- [17] 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.
- [18] Nadine Rüegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. Barc: Breed-augmented regression using classification for 3d dog reconstruction from images. *International Journal of Computer Vision*, 131(8):1964–1979, 2023. 1
- [19] Remy Sabathier, Niloy J. Mitra, and David Novotny. Animal avatars: Reconstructing animatable 3d animals from casual videos. In *ECCV*, pages 270–287, 2024. 1, 2
- [20] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: A single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 2
- [21] Samarth Sinha, Roman Shapovalov, Jeremy Reizenstein, Ignacio Rocco, Natalia Neverova, Andrea Vedaldi, and David Novotny. Common pets in 3d: Dynamic new-view synthesis of real-life deformable categories. In *CVPR*, 2023. 2
- [22] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2
- [23] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 2, 4
- [24] SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025. 2, 4
- [25] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humanerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022. 2
- [26] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 3
- [27] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 2, 4
- [28] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T. Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *CVPR*, 2021. 2
- [29] 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 2863–2873, 2022. 1, 2
- [30] Gengshan Yang, Chaoyang Wang, N. Dinesh Reddy, and Deva Ramanan. Reconstructing animatable categories from videos. In *CVPR*, 2023. 1, 2
- [31] Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent image to 3d view synthesis via geometry-aware diffusion models. *3DV*, 2023. 2
- [32] Yiyu Zhuang, Jiayi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26308–26319, 2025. 1, 2, 4
- [33] Silvia Zuffi and Michael J Black. Awol: Analysis without synthesis using language. In *ECCV*, pages 1–19. Springer, 2024. 1, 3
- [34] 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. 1, 5