MoReMouse: Monocular Reconstruction of Laboratory Mouse

Yuan Zhong*, **Jingxiang Sun***, **Zhongbin Zhang**, **Liang An**†, **Yebin Liu**†
Tsinghua University
{zhong-y19, zhangzb23}@mails.tsinghua.edu.cn,
{anliang, liuyebin}@mail.tsinghua.edu.cn,
starkjxsun@gmail.com

Abstract

Laboratory mice, particularly the C57BL/6 strain, are essential animal models in biomedical research. However, accurate 3D surface motion reconstruction of mice remains a significant challenge due to their complex non-rigid deformations, textureless fur-covered surfaces, and the lack of realistic 3D mesh models. Moreover, existing visual datasets for mice reconstruction only contain sparse viewpoints without 3D geometries. To fill the gap, we introduce MoReMouse, the first monocular dense 3D reconstruction network specifically designed for C57BL/6 mice. To achieve high-fidelity 3D reconstructions, we present three key innovations. First, we create the first high-fidelity, dense-view synthetic dataset for C57BL/6 mice by rendering a realistic, anatomically accurate Gaussian mouse avatar. Second, MoReMouse leverages a transformer-based feedforward architecture combined with triplane representation, enabling high-quality 3D surface generation from a single image, optimized for the intricacies of small animal morphology. Third, we propose geodesic-based continuous correspondence embeddings on the mouse surface, which serve as strong semantic priors, improving surface consistency and reconstruction stability, especially in highly dynamic regions like limbs and tail. Through extensive quantitative and qualitative evaluations, we demonstrate that MoReMouse significantly outperforms existing open-source methods in both accuracy and robustness.

Project page —
https://zyyw-eric.github.io/MoreMouse-webpage/
Extended version — https://arxiv.org/abs/2507.04258

1 Introduction

Laboratory mice are widely used model organisms in biomedical research, especially for disease modeling, genetic studies, and neuroscience. Among them, the C57BL/6 strain is the most commonly used inbred line due to its well-characterized genome and relevance in neurodegenerative, oncological, and immunological research. Accurate 3D motion reconstruction of C57BL/6 mice is essential for behavioral phenotyping.

Over the last decade, deep learning has revolutionized both 2D and 3D animal pose estimation, with methods such as DeepLabCut (Mathis et al. 2018), SLEAP (Pereira et al. 2022) and DANNCE (Dunn et al. 2021) becoming standard tools across diverse species including mice. However, these methods are limited to sparse keypoint tracking and ignore important 3D surface motion. To this end, template-based papers tried to build articulated mouse mesh models (Bolaños et al. 2021) and used them for mouse 3D tracking (An et al. 2023). However, the manually designed skinning weights are infeasible to represent the complex dynamic deformations of a mouse, resulting in severe self-penetration during its free movement. Moreover, mesh fitting methods require heavy and sensitive multi-view optimization, hindering their application for more convenient single-view scenarios. Template-free animal reconstruction methods such as 3D-Fauna (Li et al. 2024a) lack the knowledge of mouse-specific appearance and geometry, therefore have difficulty in recovering 3D C57BL/6 mice.

The rapid evolution of 3D generative AI has blurred the boundary between 3D reconstruction and 3D generation, with large-scale datasets such as Objaverse (Deitke et al. 2023) playing a crucial role in this trend. Fueled by such data, Large Reconstruction Models (LRMs) (Hong et al. 2023; Tochilkin et al. 2024; Tang et al. 2024) have demonstrated high efficiency and accuracy in single-view, feedforward 3D reconstruction of general objects. More recent generative frameworks, such as Trellis (Xiang et al. 2025) based on rectified flow, and CLAY (Zhang et al. 2024b) based on diffusion, further push the scalability and controllability of 3D generation by learning powerful structured latent representations. Although these models show great promise, they do not perform well on laboratory mice due to the lack of large-scale 3D mice supervision or dense-view mice images. Moreover, as these methods only consider single-image cases, no temporal semantic correspondences are ensured for long-term dynamic object reconstruction.

In this paper, we present **MoReMouse**, the first **mo**nocular dense 3D **re**construction model specifically designed for C57BL/6 laboratory **mouse**. MoReMouse is built upon a transformer and triplane representation architecture, enabling efficient, feedforward 3D reconstruction, as shown in Fig. 1. To compensate for the scarcity of dense-view 3D mouse datasets, we construct the first animatable Gaus-

*These authors contributed equally.
†Corresponding authors

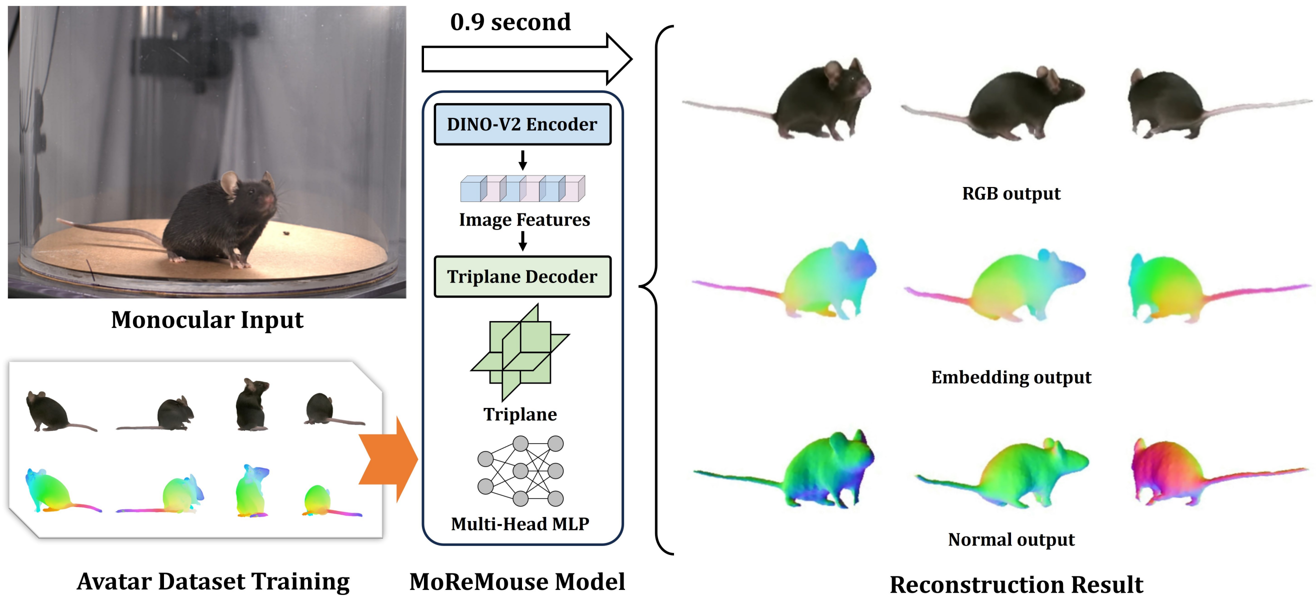

Figure 1: We present **MoReMouse**, the first monocular dense 3D reconstruction framework for laboratory mice. Given a single-view input (top-left, captured with an iPhone 15 Pro), MoReMouse predicts high-fidelity surface geometry and appearance within 0.9 seconds using a transformer-based triplane architecture (middle). To assist model training, we render a dense-view synthetic dataset by building the first Gaussian mouse avatar from sparse-view real videos (bottom-left). Our method outputs RGB renderings, semantic embeddings, and normal maps from a single image (right). Corresponding video results are provided in the supplementary material.

sian avatar of mouse (**AGAM** for short) using an articulated mouse model (Bolaños et al. 2021) and Gaussian Splatting (Kerbl et al. 2023). Inspired by recent human avatar modeling work (Li et al. 2024b), we use a 2D UV-based intermediate mouse pose representation to drive Gaussian point clouds. Thanks to the free-view rendering ability and diverse pose space of the avatar, we build a synthetic 64-view dataset, enabling effective training of the MoReMouse network.

Based on the new synthetic dataset, the framework of MoReMouse integrates multiple architectural and training enhancements inspired by LRMs. Specifically, we employ a two-stage training strategy, initially pretraining the model with triplane-NeRF (Chan et al. 2022; Mildenhall et al. 2021) representations, followed by fine-tuning with triplane-DMTet (Shen et al. 2021) for higher-fidelity reconstructions. To further improve the stability and robustness, we introduce geodesic-based feature embeddings, which enhance semantic consistency of surfaces across frames.

We evaluate MoReMouse on both synthetic and real-world datasets. The synthetic benchmark is rendered using our AGAM, while the real-world evaluation uses a 4-view video sequence of a single freely moving C57BL/6 mouse captured by Point Grey industrial cameras. Results demonstrate the superior performance of MoReMouse in monocular mouse reconstruction and novel view synthesis. As a whole, MoReMouse represents a significant step toward practical and high-fidelity monocular 3D reconstruction of C57BL/6 mice, with potential to support broader research in computational modeling and analysis of small animals. To summarize, our key contributions include:

- We introduce **MoReMouse**, the first monocular dense 3D reconstruction model tailored for laboratory mice, specifically the C57BL/6 strain, and show its superiority over existing open-source alternatives.
- We develop the first animatable Gaussian avatar model for mice and use it to generate a large-scale, dense-view synthetic dataset for training and evaluation.
- We integrate geodesic-based continuous correspondence embeddings into feedforward reconstruction, improving single-view surface stability and detail preservation.

2 Related Work

3D Representation of Mouse

Previously, most methods reconstruct mouse 3D keypoints by algebraic triangulation (Nath et al. 2019; Karashchuk et al. 2021; Han et al. 2024; Monsees et al. 2022), voxel-based triangulation (Dunn et al. 2021) or even by optical markers (Marshall et al. 2021). However, 3D keypoints provide only sparse geometric structures, neglecting essential dense surface information. To represent the mouse surface, an articulated mesh model was proposed (Bolaños et al. 2021) and subsequently used for multi-view mesh fitting (An et al. 2023). ArMo (Bohnslav et al. 2023) presents a similar mouse mesh model, but lacks true anatomical bones. For rats, which are similar to mice in shape, similar mesh models are built (Aldarondo et al. 2024; Klibaite et al. 2025). All these mesh models are limited to the manually designed skinning weights, which could not capture the realistic dynamic surface motion of mice.

Monocular Reconstruction of Animals

Since the proposal of the SMAL model which represents several quadrupedal species (Zuffi et al. 2017), predicting the pose and shape parameters of SMAL from images has dominated the monocular reconstruction of animals (Biggs et al. 2020; Xu et al. 2023; Rueegg et al. 2022; Lyu et al. 2025). Although they perform well for quadrupeds, they cannot reconstruct rodents due to the limited shape space of SMAL. The AWOL model (Zuffi and Black 2024) extends SMAL to represent rodents using CLIP (Radford et al. 2021), yet its pose and shape realism are still limited. Some recent papers try to build animal avatars from videos by augmenting the skinning weights or adding additional deformations, such as BANMo (Yang et al. 2022), Gart (Lei et al. 2024) and AnimalAvatars (Sabathier, Mitra, and Novotny 2024). However, these methods all require a heavy optimization process during inference, and rely on dense pose estimation or accurate SMAL parameters for initialization. Template-free methods such as MagicPony (Wu et al. 2023a) and 3D Fauna (Li et al. 2024a) leverage collections of single-view images of deformable species captured in the wild to jointly learn canonical models and pose regressors. However, these approaches do not generalize well to mice due to domain and geometry gaps.

Large Reconstruction Models

Large reconstruction models have recently achieved substantial progress in generating high-quality 3D content (Hong et al. 2023; Li et al. 2023; Zou et al. 2024; Zhang et al. 2024a; Sun et al. 2024). These models are typically trained in an end-to-end fashion to predict implicit 3D representations, often through triplane-based neural radiance fields. Recent methods improve quality and efficiency by learning generative models over compact latent codes derived via VAE-like encoders (Hu et al. 2025). Some focus purely on shape (Zhang et al. 2023; Li et al. 2025), while others incorporate appearance (Gupta et al. 2023; Xiong et al. 2024). CLAY (Zhang et al. 2024b) presents a unified system for geometry and PBR texture generation. Trellis (Xiang et al. 2025) introduces structured latent codes via rectified flow transformers, jointly modeling geometry and appearance for versatile decoding into NeRFs, Gaussians, and meshes. However, these methods generally depend on access to 3D supervision (Deitke et al. 2023; Wu et al. 2023b), such as ground-truth shapes or high-quality multi-view data, during training. The data scarcity is one of the main challenges for training an LRM for laboratory mice.

3 Method

We aim to feedforwardly reconstruct the dense surface geometry of a laboratory mouse from a single image, focusing on the widely used C57BL/6 strain.

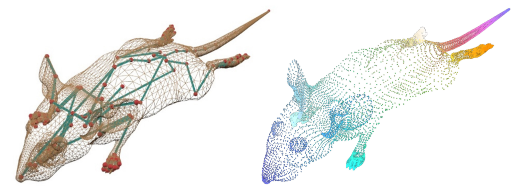

Figure 2: Left: the articulated mouse mesh we used. Right: the color-coded geodesic feature embedding, promoting surface separability and reconstruction consistency.

Synthetic Dataset Using Gaussian Avatar

Mouse Parametric Model. As shown in Fig 2, we modify the mouse model from (Bolaños et al. 2021) following the approach of (An et al. 2023), where redundant mesh vertices (teeth and tongue) are removed. The model consists of $N_J = 140$ articulated joints and $N_V = 13059$ vertices, and conforms to Linear Blend Skinning (LBS) for posing. Poses of this refined model serve as the avatar control signals. The whole parameter set is denoted as Ψ, including joint rotations $\theta \in \mathbb{R}^{3N_J}$, global translation $t \in \mathbb{R}^3$, global rotation $r \in \mathbb{R}^3$, global scale $s \in \mathbb{R}$ and bone length deformation parameters $B \in \mathbb{R}^{21}$ (please refer to (An et al. 2023) for more details). We decompose Ψ into local parameters $\Psi_l = \{\theta, B\}$ and global parameters $\Psi_g = \{r, t, s\}$ for the sake of simplicity.

Gaussian Mouse Avatar. Gaussian point clouds (Kerbl et al. 2023) use an explicit parameterization for 3D representation. Each Gaussian point is characterized by a mean position $\mu \in \mathbb{R}^3$ and an anisotropic covariance matrix Σ:

$$G(x) = e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}. \tag{1}$$

To facilitate optimization, the covariance matrix is reformulated as $\Sigma = RSS^T R^T$, where S is a diagonal scaling matrix controlled by a three-dimensional vector $s \in \mathbb{R}^3$, and the rotation matrix R is represented by a normalized quaternion $q \in \mathbb{R}^4$. Additionally, each Gaussian encodes a color $c \in \mathbb{R}^3$ and an opacity $o \in \mathbb{R}$. During rendering, Gaussians within the pixel projection region are sorted by depth and composited using alpha blending.

The pipeline of our AGAM is shown in Fig. 3. Our key idea is to parameterize the mouse UV texture, mapping each valid texel in the UV space to a corresponding Gaussian point, where the texture color represents the Gaussian's position. To achieve global-pose invariant control of AGAM, we set $\Psi_g = 0$ and drive the mesh model using only Ψ_l, resulting in canonical vertex coordinates V_{Ψ_l}. The UV-mapped positions of these vertices are used to interpolate Gaussian locations via barycentric coordinates. This defines the baseline Gaussian position map μ_{Ψ_l}:

$$\mu_{\Psi_l} = \mathcal{D}_{\Psi_l}(\mu_0), \tag{2}$$

where μ_0 is the UV coordinate map in the neutral pose, and $\mathcal{D}_{\Psi_l}$ represents the LBS deformation mapping induced by the skeletal model.

The core network of AGAM is a set of StyleUNets (Wang et al. 2023) that predict the position offset map $\Delta\mu_{\Psi_l}$ and

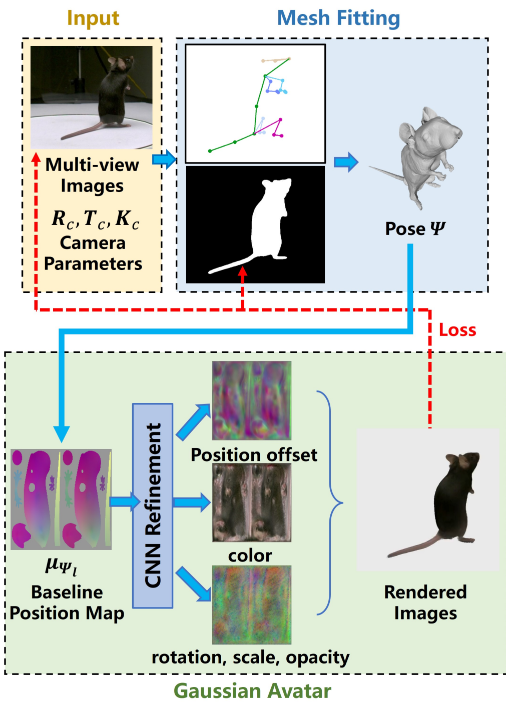

Figure 3: Pipeline of the Gaussian Mouse Avatar Generation. The process begins with Mesh Fitting, followed by Gaussian Avatar training. "CNN" here is StyleUNet.

other Gaussian attributes c_{Ψ_l}, o_{Ψ_l}, q_{Ψ_l}, and s_{Ψ_l} conditioned on the UV coordinate map μ_{Ψ_l}. The Gaussian positions and rotations are deformed jointly via the skeletal deformation operator by $\mu'_{\Psi_l} = \mathcal{D}_{\Psi_l}(\mu_0 + \Delta\mu_{\Psi_l})$ and $q'_{\Psi_l} = \mathcal{D}_{\Psi_l}(q_{\Psi_l})$, respectively. Then the Gaussian Splatting rendering function generates the view-dependent rendered image I_c^{rend}:

$$I_c^{\text{rend}} = GS(R_c, T_c, K_c, \mu'_{\Psi_l}, c_{\Psi_l}, q'_{\Psi_l}, s_{\Psi_l}), \tag{3}$$

where $GS(\cdot)$ denotes the Gaussian splatting renderer, and R_c, T_c, K_c are camera parameters.

Avatar Training. We define loss functions between the real image I_c and the rendered image I_c^{rend} as:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_1(I_c - I_c^{\text{rend}}) + \lambda_{\text{SSIM}}(1 - \mathcal{L}_{\text{SSIM}}(I_c, I_c^{\text{rend}})) \\ + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}(I_c, I_c^{\text{rend}}),\end{aligned} \tag{4}$$

where $\mathcal{L}_{\text{SSIM}}$ is structural similarity loss, $\mathcal{L}_{\text{LPIPS}}$ is perceptual similarity loss, $\lambda_{\text{SSIM}} = 0.2$ and $\lambda_{\text{LPIPS}} = 0.1$. We use Total Variation (TV) loss to enforce point smoothness. See the supplementary material for further details.

Since C57BL/6 mice usually show similar texture, we only use a 6-view mouse video sequence "markerless_mouse_1" proposed by (Dunn et al. 2021) for avatar training. This sequence captures 18,000 frames of a single freely moving mouse in an open field at 1152×1024@100fps, where the mouse mesh models are borrowed from (An et al. 2023). Note that we only use 800 frames for training, which are uniformly sampled from the first 8000 frames of "markerless_mouse_1". Avatar training takes 400k steps.

Dense-View Dataset. We drive AGAM using the poses from all frames of the "markerless_mouse_1" video sequence, and render evenly distributed 64-view synthetic images. Among the full 18,000 frames, we use the first 6,000 frames for training and the last 6,000 frames for testing. Note that the 800 frames used for avatar creation all belong to the first 6,000 frames, therefore the poses in the test set are unseen. To construct the training set, we sample two sets of 64 viewpoints per frame, rendering 12,000 multi-view scenes in total. Please refer to the supplementary file for more information about view selection and data samples.

MoReMouse Architecture

Transformer and Triplane Representation. Similar to LRM (Hong et al. 2023), MoReMouse adopts a transformer and triplane-based architecture, designed for monocular 3D reconstruction. As illustrated in Fig. 4, the model consists of three key components: an image encoder, an image-to-triplane decoder, and a triplane-based geometric representation that integrates both NeRF (Mildenhall et al. 2021) and DMTet (Shen et al. 2021) for implicit and explicit geometry modeling.

We use a pretrained DINOv2 encoder to extract latent features from the input image. These features are decoded via a transformer with cross-attention to generate a triplane representation, which encodes color, density, and semantic attributes for any 3D query point.

During the initial training stage, we employ NeRF for volumetric rendering without introducing explicit geometry, enabling smooth gradient propagation through the entire density field. After convergence, we switch to DMTet to obtain an explicit surface representation for fine-level geometric supervision.

The fine-scale structures of mice, particularly in the limbs and tail, present challenges for high-fidelity reconstruction in a feedforward manner. To address this, we modify the LRM framework by increasing the triplane resolution to 64 to improve detail preservation. Additional architectural details are provided in the supplementary materials.

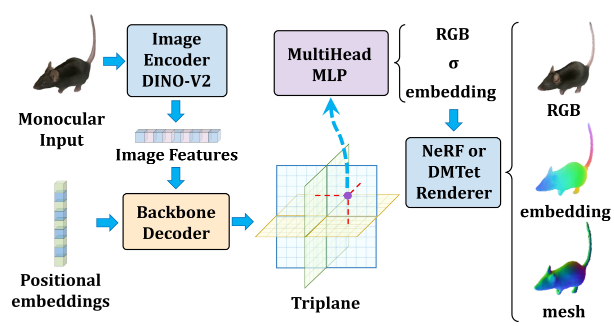

Figure 4: Overview of the MoReMouse architecture. Given a single image, a DINOv2 encoder and transformer decoder generate triplane features, which are queried to produce color, density, and embeddings. Rendering is performed via NeRF or DMTet for volumetric or surface outputs.

Feature Embedding for Surface Consistency. Mouse reconstruction is challenging due to black fur with weak texture cues and fine-scale surface details. Inspired by Animatable Gaussians (Li et al. 2024b), we incorporate feature embeddings computed from geodesic distances (Shamai and Kimmel 2017) into the surface representation. These embeddings are used as additional color channels during training, serving as semantic priors to enhance reconstruction stability.

As shown in Fig. 2, given the set of mouse model vertices $V \in \mathbb{R}^{N_V \times 3}$, we compute the pairwise geodesic distances between vertices to obtain the geodesic distance matrix:

$$G_V \in \mathbb{R}^{N_V \times N_V}, \quad G_V(i,j) = d_{\text{geo}}(V_i, V_j), \tag{5}$$

where $d_{\text{geo}}(V_i, V_j)$ represents the geodesic distance between vertices V_i and V_j.

To learn a continuous correspondence embedding that captures the intrinsic structure of the mouse surface, we define the embedding matrix $E \in \mathbb{R}^{N_V \times 3}$ and optimize it by minimizing the L_2 loss between the embedding-derived distance matrix G_E and the geodesic distance matrix G_V:

$$\mathcal{L}_{\text{geo}} = \sum_{i,j} \| G_E(i,j) - G_V(i,j) \|_2^2 , \tag{6}$$

where $G_E(i,j) = \|E_i - E_j\|_2$ represents the Euclidean distance in the embedding space.

However, directly using these embeddings as color channels is suboptimal. The three-dimensional embeddings often form an elongated mouse-like shape in feature space, failing to fully utilize the available color space. Additionally, large dark or bright regions appear, lacking sufficient contrast against the black or white backgrounds used during rendering. To address this, we perform Principal Component Analysis (PCA) on the embeddings, selecting the first two principal components as the hue (H) and saturation (S) values in the HSV color space, while keeping the value (V) channel fixed at 1:

$$H = \text{PCA}_1(E), \quad S = \text{PCA}_2(E), \quad V = 1. \tag{7}$$

This transformation improves the separability of the embeddings in the rendered images.

Training and Optimization. During training, two randomly selected viewpoints from each scene serve as the input-output pair. To mitigate spatial ambiguity, we fix the mouse position at the scene origin. In preprocessing, we also apply translation and scaling transformations to the input images, ensuring that the mouse centroid is aligned with the image center and normalized to a consistent scale. Experimental results show that MoReMouse generalizes well from training on fixed perspectives to varying viewpoints in real-world scenarios.

We employ a two-stage training strategy, where NeRF-based volumetric rendering is first trained for 60 epochs to ensure proper density field convergence, followed by DMTet-based explicit geometry modeling for an additional 100 epochs to refine surface details. Details of the training loss terms and hyperparameters are provided in the supplementary material.

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓
synthetic	**Ours**	22.027	0.9660	0.05279
	TripoSR	13.673	0.8032	0.18255
	Triplane-GS	18.049	0.9268	0.10151
	LGM	14.460	0.8805	0.13274
	InstantMesh	15.821	0.8987	0.11312
real	**Ours**	18.422	0.9478	0.08674
	TripoSR	11.518	0.8114	0.19672
	Triplane-GS	16.789	0.9298	0.11002
	LGM	15.215	0.9197	0.12838
	InstantMesh	15.631	0.9175	0.11339

Table 1: Quantitative results on both synthetic and real-world datasets. MoReMouse achieves the best performance across all metrics in both domains.

4 Experiments

Experimental Setup

We evaluate MoReMouse on two types of testing datasets: a synthetic benchmark and a real-world captured dataset.

Synthetic Evaluation Data. We use the last 6,000 frames from the "markerless_mouse_1" sequence to generate a synthetic multi-view dataset consisting of 4 orthogonal viewpoints. These views are rendered using our Gaussian avatar pipeline and serve as a controlled evaluation set where ground-truth camera parameters and rendering conditions are known.

Real Captured Data. To assess performance under realistic conditions, we collect a real-world multi-view dataset using Point Grey industrial cameras. This dataset contains 5,400 frames of freely moving mice captured from four calibrated viewpoints. More details about camera setup and calibration are provided in the supplementary materials.

Baselines. Due to the lack of monocular 3D reconstruction methods specifically designed for mice, we compare MoReMouse against four general-purpose single-view reconstruction baselines:

- **TripoSR** (Tochilkin et al. 2024): A state-of-the-art open-source Large Reconstruction Model (LRM) known for strong single-view 3D generation performance.
- **Triplane-GS** (Zou et al. 2024): A LRM variant based on Gaussian representations.
- **LGM** (Tang et al. 2024): A multi-view diffusion-based method that reconstructs 3D Gaussian point clouds.
- **InstantMesh** (Xu et al. 2024): A mesh-based 3D reconstruction method using multi-view diffusion priors.

All methods are evaluated on the task of novel view synthesis, since obtaining ground-truth 3D geometry for mouse

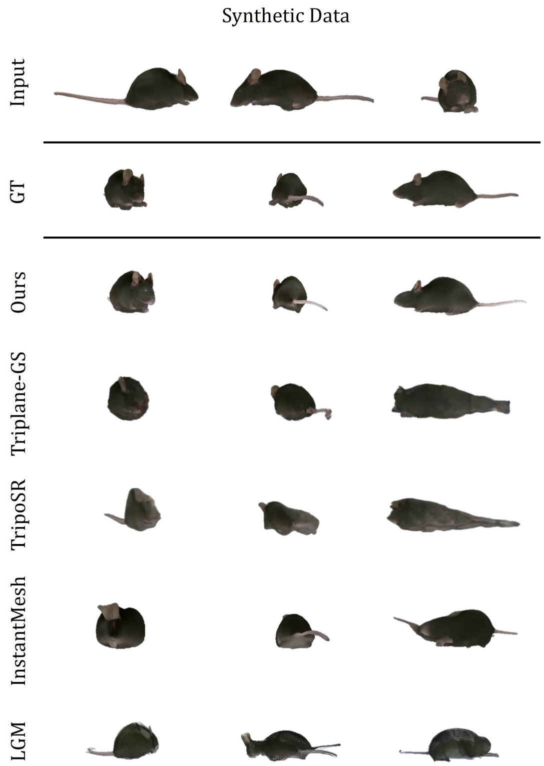

Figure 5: Qualitative comparison of novel view synthesis results on synthetic data. MoReMouse produces coherent and anatomically plausible reconstructions.

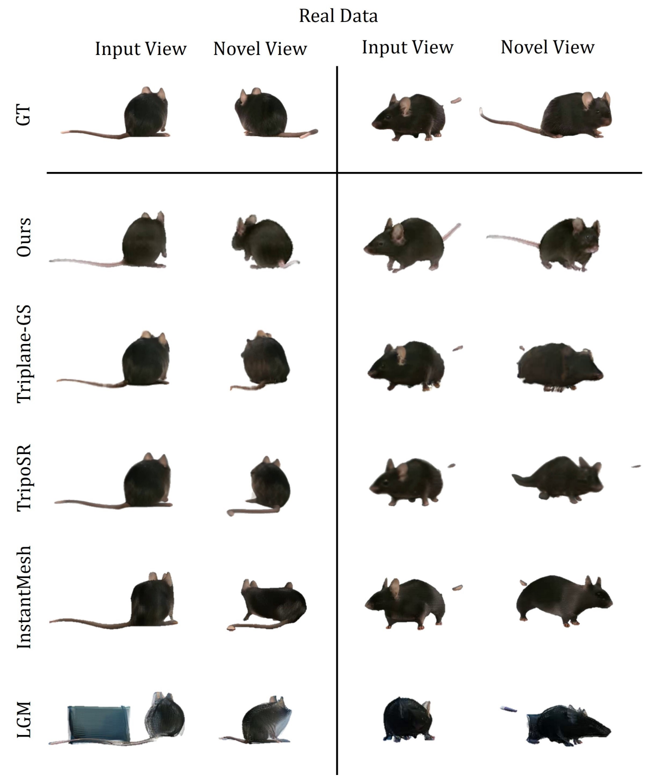

Figure 6: Qualitative comparison on real data. MoReMouse maintains consistent geometry and appearance, while baselines often show pose and shape artifacts. Occasional tail discontinuities arise from imperfect Track-Anything masks under reflections or occlusions; our model remains robust and recovers plausible tail shapes.

subjects is impractical. Additionally, we include surface geometry metrics (IoU against multi-view visual hulls) in the supplementary material. We use Track-Anything (Yang et al. 2023) to remove background regions in real videos.

Experimental Results

Quantitative Results. We report the evaluation results on the synthetic and real-world datasets in Table 1.

Across all three metrics for novel view synthesis, MoReMouse consistently outperforms the baseline methods, demonstrating its ability to effectively infer plausible mouse geometry and pose from a single image. We also observe that the quantitative metrics on the real-world dataset are generally lower than on the synthetic dataset, especially for PSNR. This is largely due to the camera-mouse configuration: in the real dataset, the cameras are fixed while the mouse moves, whereas MoReMouse and all baselines assume that the mouse is centered and aligned across input-output views. This mismatch introduces pixel-level alignment errors, reducing PSNR. However, the perceptual metrics SSIM and LPIPS remain high, indicating that MoReMouse still reconstructs plausible and correct mouse poses and shapes. This confirms that models trained on synthetic data can generalize to real-world mouse motion capture scenarios. Full tables of IoU metrics, expanded baseline comparisons, and runtime/memory usage are provided in the supplementary materials.

Qualitative Results. To visually compare MoReMouse with baseline methods, we select three representative frames from both the synthetic and real-world sequences. As shown in Fig. 5 and Fig. 6, MoReMouse generates mice with coherent appearance and physically plausible postures, accurately capturing tail and limb movements. These results remain stable even though each frame is inferred independently, showing smooth motion continuity. In contrast, baseline methods often exhibit structural distortions in novel views and inconsistent pose transitions across frames.

Ablation Study

Comparison with Fine-tuned TripoSR. To further validate the benefits of our architectural design, we fine-tune TripoSR on our mouse dataset for 120 epochs with a learning rate of 1×10^{-6}. As shown in Table 2 and Fig. 7, the fine-tuned TripoSR shows noticeable improvement over its original version but still underperforms MoReMouse across most metrics. This indicates that, beyond data adaptation, our architectural choices—such as triplane encoding and geodesic embeddings—contribute significantly to reconstruction quality. We also fine-tune Triplane-GS and re-

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓
synthetic	Ours	22.027	0.9660	0.05279
	TripoSR-Tuned	21.996	0.9676	0.06333
real	Ours	18.422	0.9478	0.08674
	TripoSR-Tuned	18.162	0.9461	0.09354

Table 2: Comparison between MoReMouse and TripoSR fine-tuned on mouse data.

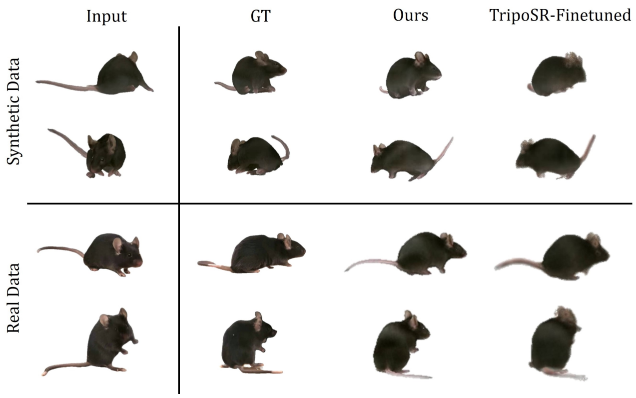

Figure 7: Qualitative comparison between our MoReMouse method and a fine-tuned TripoSR baseline. MoReMouse produces sharper and more anatomically coherent results.

port its performance in the supplementary, which reveals reconstruction artifacts particularly around limbs and the tail.

Effect of Feature Embedding. We evaluate the effectiveness of the geodesic feature embedding module by training a variant of MoReMouse without this component. As shown in Table 3 and visualized in Fig. 8, removing the embedding leads to noticeably degraded reconstruction quality, particularly around fine-scale structures.

Without embedding supervision, the model struggles to maintain surface consistency in anatomically delicate regions such as limbs and tails. As evident in the first, third, and fourth rows of Fig. 8, the full model produces sharper and more accurate reconstructions of the paws and extremities. In the second row, the embedding-based model also better preserves subtle texture details around the snout. These observations support our design choice to use surface-based semantic priors. The geodesic embedding encourages coherent geometry and finer detail recovery.

5 Discussion

Summary. In this paper, we present **MoReMouse**, the first monocular dense 3D reconstruction model tailored for laboratory mice. Unlike existing single-view 3D methods designed for general objects or humans, MoReMouse addresses the unique challenges posed by small animals, including non-rigid deformation, low surface texture, and limited training data. Our approach integrates transformer-

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓
synthetic	full model	22.027	0.9660	0.05279
	w/o embedding	21.805	0.9655	0.05467
real	full model	18.422	0.9478	0.08674
	w/o embedding	18.250	0.9469	0.08767

Table 3: Ablation study on the effect of feature embedding. Removing embedding results in lower reconstruction quality, especially in perceptual similarity.

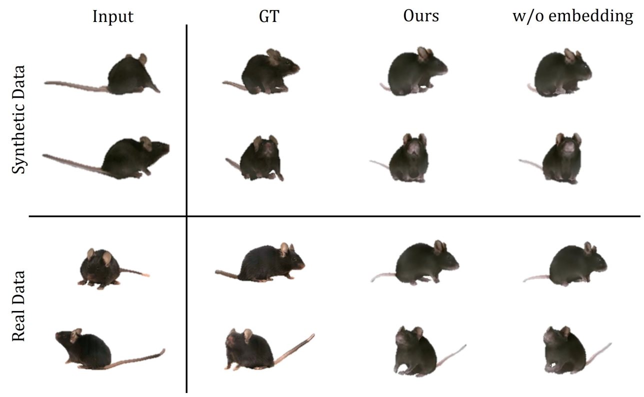

Figure 8: Qualitative comparison with and without geodesic feature embedding. The full model produces sharper and more consistent geometry, especially around limbs (rows 1, 3, 4) and snout texture (row 2). Embeddings help guide fine-grained surface details and improve anatomical accuracy.

based triplane representations, a Gaussian avatar modeling pipeline, and geodesic surface feature embeddings to achieve robust and high-fidelity reconstructions from a single image. We demonstrate strong performance across both synthetic and real datasets, outperforming state-of-the-art reconstruction baselines in terms of visual quality and perceptual consistency. Ablation studies further validate the architectural contributions of our model beyond dataset-specific priors. We hope that MoReMouse can serve as a foundation for future research in computational ethology and animal modeling, promoting more precise, efficient, and scalable behavioral analysis tools for biomedical research.

Limitations. Although MoReMouse offers a promising solution to monocular mouse reconstruction, several limitations remain that warrant future investigation. (1) *Limited training diversity.* Our training data is derived from a single video sequence of a single mouse under consistent lighting and appearance. This limits the generalization of the model to other mice with varying fur color, size, or under different lighting conditions. (2) *Pose tracking in global coordinates.* The current system assumes that the mouse is always centered in a canonical scene space. This design limits our ability to track full 3D poses in a real-world coordinate system.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62125107) and the Shuimu Tsinghua Scholar Program (2024SM324).

References

Aldarondo, D.; Merel, J.; Marshall, J. D.; Hasenclever, L.; Klibaite, U.; Gellis, A.; Tassa, Y.; Wayne, G.; Botvinick, M.; and Ölveczky, B. P. 2024. A virtual rodent predicts the structure of neural activity across behaviours. *Nature*, 632(8025): 594–602.

An, L.; Ren, J.; Yu, T.; Hai, T.; Jia, Y.; and Liu, Y. 2023. Three-dimensional surface motion capture of multiple freely moving pigs using MAMMAL. *Nature Communications*, 14(1): 7727.

Biggs, B.; Boyne, O.; Charles, J.; Fitzgibbon, A.; and Cipolla, R. 2020. Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop. In *European Conference on Computer Vision*, 195–211. Springer.

Bohnslav, J. P.; Osman, M. A. M.; Jaggi, A.; Soares, S.; Weinreb, C.; Datta, S. R.; and Harvey, C. D. 2023. ArMo: An Articulated Mesh Approach for Mouse 3D Reconstruction. *bioRxiv*, 2023–02.

Bolaños, L. A.; Xiao, D.; Ford, N. L.; LeDue, J. M.; Gupta, P. K.; Doebeli, C.; Hu, H.; Rhodin, H.; and Murphy, T. H. 2021. A three-dimensional virtual mouse generates synthetic training data for behavioral analysis. *Nature Methods*, 18(4): 378–381.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16123–16133.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13142–13153.

Dunn, T. W.; Marshall, J. D.; Severson, K. S.; Aldarondo, D. E.; Hildebrand, D. G.; Chettih, S. N.; Wang, W. L.; Gellis, A. J.; Carlson, D. E.; Aronov, D.; et al. 2021. Geometric deep learning enables 3D kinematic profiling across species and environments. *Nature methods*, 18(5): 564–573.

Gupta, A.; Xiong, W.; Nie, Y.; Jones, I.; and Oğuz, B. 2023. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*.

Han, Y.; Chen, K.; Wang, Y.; Liu, W.; Wang, Z.; Wang, X.; Han, C.; Liao, J.; Huang, K.; Cai, S.; et al. 2024. Multi-animal 3D social pose estimation, identification and behaviour embedding with a few-shot learning framework. *Nature Machine Intelligence*, 1–14.

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; and Tan, H. 2023. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*.

Hu, Y.; Wang, L.; Liu, X.; Chen, L.-H.; Guo, Y.; Shi, Y.; Liu, C.; Rao, A.; Wang, Z.; and Xiong, H. 2025. Simulating the real world: A unified survey of multimodal generative models. *arXiv preprint arXiv:2503.04641*.

Karashchuk, P.; Rupp, K. L.; Dickinson, E. S.; Walling-Bell, S.; Sanders, E.; Azim, E.; Brunton, B. W.; and Tuthill, J. C. 2021. Anipose: a toolkit for robust markerless 3D pose estimation. *Cell Reports*, 36(13).

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. *TOG*, 42(4): 1–14.

Klibaite, U.; Li, T.; Aldarondo, D.; Akoad, J. F.; Ölveczky, B. P.; and Dunn, T. W. 2025. Mapping the landscape of social behavior. *Cell*, 188(8): 2249–2266.

Lei, J.; Wang, Y.; Pavlakos, G.; Liu, L.; and Daniilidis, K. 2024. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19876–19887.

Li, J.; Tan, H.; Zhang, K.; Xu, Z.; Luan, F.; Xu, Y.; Hong, Y.; Sunkavalli, K.; Shakhnarovich, G.; and Bi, S. 2023. Instant3D: Fast Text-to-3D with Sparse-View Generation and Large Reconstruction Model. *ArXiv*, abs/2311.06214.

Li, W.; Liu, J.; Yan, H.; Chen, R.; Liang, Y.; Chen, X.; Tan, P.; and Long, X. 2025. CraftsMan3D: High-fidelity Mesh Generation with 3D Native Diffusion and Interactive Geometry Refiner. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5307–5317.

Li, Z.; Litvak, D.; Li, R.; Zhang, Y.; Jakab, T.; Rupprecht, C.; Wu, S.; Vedaldi, A.; and Wu, J. 2024a. Learning the 3d fauna of the web. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9752–9762.

Li, Z.; Zheng, Z.; Wang, L.; and Liu, Y. 2024b. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19711–19722.

Lyu, J.; Zhu, T.; Gu, Y.; Lin, L.; Cheng, P.; Liu, Y.; Tang, X.; and An, L. 2025. AniMer: Animal Pose and Shape Estimation Using Family Aware Transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17486–17496.

Marshall, J. D.; Klibaite, U.; Gellis, A.; Aldarondo, D. E.; Ölveczky, B. P.; and Dunn, T. W. 2021. The pair-r24m dataset for multi-animal 3d pose estimation. *bioRxiv*, 2021–11.

Mathis, A.; Mamidanna, P.; Cury, K. M.; Abe, T.; Murthy, V. N.; Mathis, M. W.; and Bethge, M. 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9): 1281–1289.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Monsees, A.; Voit, K.-M.; Wallace, D. J.; Sawinski, J.; Charyasz, E.; Scheffler, K.; Macke, J. H.; and Kerr, J. N. 2022. Estimation of skeletal kinematics in freely moving rodents. *Nature Methods*, 19(11): 1500–1509.

Nath, T.; Mathis, A.; Chen, A. C.; Patel, A.; Bethge, M.; and Mathis, M. W. 2019. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature Protocols*, 14(7): 2152–2176.

Pereira, T. D.; Tabris, N.; Matsliah, A.; Turner, D. M.; Li, J.; Ravindranath, S.; Papadoyannis, E. S.; Normand, E.; Deutsch, D. S.; Wang, Z. Y.; et al. 2022. SLEAP: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4): 486–495.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rueegg, N.; Zuffi, S.; Schindler, K.; and Black, M. J. 2022. Barc: Learning to regress 3d dog shape from images by exploiting breed information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3876–3884.

Sabathier, R.; Mitra, N. J.; and Novotny, D. 2024. Animal avatars: Reconstructing animatable 3D animals from casual videos. In *European Conference on Computer Vision*, 270–287. Springer.

Shamai, G.; and Kimmel, R. 2017. Geodesic distance descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6410–6418.

Shen, T.; Gao, J.; Yin, K.; Liu, M.-Y.; and Fidler, S. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34: 6087–6101.

Sun, J.; Peng, C.; Shao, R.; Guo, Y.-C.; Zhao, X.; Li, Y.; Cao, Y.; Zhang, B.; and Liu, Y. 2024. DreamCraft3D++: Efficient Hierarchical 3D Generation with Multi-Plane Reconstruction Model. *arXiv preprint arXiv:2410.12928*.

Tang, J.; Chen, Z.; Chen, X.; Wang, T.; Zeng, G.; and Liu, Z. 2024. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, 1–18. Springer.

Tochilkin, D.; Pankratz, D.; Liu, Z.; Huang, Z.; Letts, A.; Li, Y.; Liang, D.; Laforte, C.; Jampani, V.; and Cao, Y.-P. 2024. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*.

Wang, L.; Zhao, X.; Sun, J.; Zhang, Y.; Zhang, H.; Yu, T.; and Liu, Y. 2023. Styleavatar: Real-time photo-realistic portrait avatar from a single video. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–10.

Wu, S.; Li, R.; Jakab, T.; Rupprecht, C.; and Vedaldi, A. 2023a. Magicpony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8792–8802.

Wu, T.; Zhang, J.; Fu, X.; Wang, Y.; Ren, J.; Pan, L.; Wu, W.; Yang, L.; Wang, J.; Qian, C.; et al. 2023b. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 803–814.

Xiang, J.; Lv, Z.; Xu, S.; Deng, Y.; Wang, R.; Zhang, B.; Chen, D.; Tong, X.; and Yang, J. 2025. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21469–21480.

Xiong, B.; Wei, S.-T.; Zheng, X.-Y.; Cao, Y.-P.; Lian, Z.; and Wang, P.-S. 2024. Octfusion: Octree-based diffusion models for 3d shape generation. *arXiv preprint arXiv:2408.14732*.

Xu, J.; Cheng, W.; Gao, Y.; Wang, X.; Gao, S.; and Shan, Y. 2024. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*.

Xu, J.; Zhang, Y.; Peng, J.; Ma, W.; Jesslen, A.; Ji, P.; Hu, Q.; Zhang, J.; Liu, Q.; Wang, J.; et al. 2023. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9099–9109.

Yang, G.; Vo, M.; Neverova, N.; Ramanan, D.; Vedaldi, A.; and Joo, H. 2022. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2863–2873.

Yang, J.; Gao, M.; Li, Z.; Gao, S.; Wang, F.; and Zheng, F. 2023. Track Anything: Segment Anything Meets Videos. arXiv:2304.11968.

Zhang, B.; Tang, J.; Niessner, M.; and Wonka, P. 2023. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)*, 42(4): 1–16.

Zhang, K.; Bi, S.; Tan, H.; Xiangli, Y.; Zhao, N.; Sunkavalli, K.; and Xu, Z. 2024a. GS-LRM: Large Reconstruction Model for 3D Gaussian Splatting. *arXiv*.

Zhang, L.; Wang, Z.; Zhang, Q.; Qiu, Q.; Pang, A.; Jiang, H.; Yang, W.; Xu, L.; and Yu, J. 2024b. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. *ACM Transactions on Graphics (TOG)*, 43(4): 1–20.

Zou, Z.-X.; Yu, Z.; Guo, Y.-C.; Li, Y.; Liang, D.; Cao, Y.-P.; and Zhang, S.-H. 2024. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10324–10335.

Zuffi, S.; and Black, M. J. 2024. AWOL: Analysis WithOut Synthesis Using Language. In *European Conference on Computer Vision*, 1–19. Springer.

Zuffi, S.; Kanazawa, A.; Jacobs, D. W.; and Black, M. J. 2017. 3D menagerie: Modeling the 3D shape and pose of animals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6365–6373.