International Journal of Computer Vision
https://doi.org/10.1007/s11263-025-02485-5

DogRecon: Canine Prior-Guided Animatable 3D Gaussian Dog Reconstruction From A Single Image

Gyeongsu Cho[1] · Changwoo Kang[1] · Donghyeon Soon[2] · Kyungdon Joo[1]

Received: 11 September 2024 / Accepted: 13 May 2025

Abstract

We tackle animatable 3D dog reconstruction from a single image, noting the overlooked potential of animals. Particularly, we focus on dogs, emphasizing their intrinsic characteristics that complicate 3D observation. First, the considerable variation in shapes across breeds presents a complexity for modeling. Additionally, the nature of quadrupeds leads to frequent joint occlusions compared to humans. These challenges make 3D reconstruction from 2D observations difficult, and it becomes dramatically harder when constrained to a single image. To address these challenges, our insight is to combine the acquisition of appearance from generative models, without additional data, with geometric guidance provided by a parametric representation, aiming to achieve complete geometry. To this end, we present `DogRecon`, our framework consists of two key components: Canine-centric novel view synthesis with canine prior for multi-view generation of dog and a reliable sampling weight strategy with Gaussian Splatting for animatable 3D dog reconstruction. Extensive experiments on the GART, DFA, and internet-sourced datasets confirm our framework has state-of-the-art performance in image-to-3D generation and comparable performance in animatable 3D reconstruction. Additionally, we demonstrate novel pose animation and text-to-3D dog reconstruction as applications. Project page: https://vision3d-lab.github.io/dogrecon/

Keywords Animal reconstruction · Gaussian Splatting · Dogs · Novel View Synthesis

1 Introduction

Accurate digitization of real-world components (*e.g.,* animals and humans) into 3D models is vital for creating highly realistic and engaging AR/VR applications. Traditionally, reconstructing photorealistic 3D models requires multi-view video (Peng et al., 2021a; Liu et al., 2021; Kwon et al., 2021) or scanning (Bagautdinov et al., 2021; Saito et al., 2021), both of which can be costly and complex, involving specialized equipment and technical expertise. Recently, there have been various attempts to reconstruct and even animate photorealistic models from monocular images (Lei et al., 2024; Weng et al., 2022), and even single-image situations (Huang et al., 2023; AlBahar et al., 2023), which are more challenging. Although these approaches are successful in creating 3D models and animations, their focus is primarily on human subjects (Peng et al., 2021b; Jiang et al., 2022; Huang et al., 2023).

Without a doubt, human-specific approaches are important, but humans are only a tiny part of the actual biodiversity. Therefore, we need to be capable of reconstructing the shape and pose of 3D models of broader species to eventually represent the natural world (Li et al., 2024). In this context, 3D reconstruction research is being conducted on animals (Yang et al., 2022; Li et al., 2024; Wu et al., 2023). Among these animals, dogs stand out as a quadrupedal and human-friendly species that have continued to be a subject of study in recent years (Rüegg et al., 2023a, b; Biggs et al., 2020).

However, accurately deriving the shape and pose of a 3D dog model from 2D observations remains a significant

Communicated by Shangzhe Wu.

✉ Kyungdon Joo
kyungdon@unist.ac.kr

Gyeongsu Cho
threedv@unist.ac.kr

Changwoo Kang
kangchangwoo@unist.ac.kr

Donghyeon Soon
dhsoon@dgist.ac.kr

[1] Artificial Intelligence Graduate School, UNIST, Ulsan, South Korea

[2] Department of Computer Science, DGIST, Daegu, South Korea

Published online: 02 June 2025

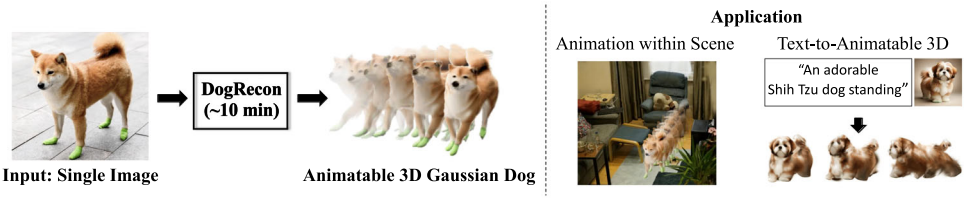

Fig. 1 Teaser of DogRecon. *Left:* The proposed DogRecon takes a single image as input and reconstructs 3D Gaussian, including texture and shape. *Right:* With the obtained 3D dogs, we can retarget motion to an existing video or even create new animations within pre-trained scenes by editing. In addition, DogRecon seamlessly applies to text-to-animatable 3D dogs.

challenge. Firstly, the wide range of breed-specific characteristics, such as the length of each leg, tail, and textures, makes the task complex. Secondly, since dogs are quadrupedal animals, their joints are often occluded even when in a standing posture. The diversity in shapes and frequent occlusion of joints create significant challenges in accurately inferring the shape and pose of dogs without sufficient 3D observations. Furthermore, we assume we use only a single image as input in this work. This setting makes reconstructing the 3D dog more challenging.

In this work, we propose a new framework called DogRecon that generates an animatable 3D dog model from a single image (see Fig. 1). Our key idea is that if we cannot predict the correct 3D model at various viewpoints, we can utilize a well-predicted 3D model as a strong cue for other challenging viewpoints. Concretely, we predict a 3D model from an input image of a dog. Then, we explicitly rotate the predicted 3D model and use it as the initial 3D model of the desired view. Our DogRecon framework consists of two modules. Firstly, we propose a canine-centric Novel View Synthesis (NVS) module that can generate multi-view images specialized to the canine geometry. The NVS module produces canine-centric multi-view images that are geometrically aligned with each other by using mask guidance within a latent diffusion model. Secondly, we propose a weighting strategy to address unreliable generated images in 3D reconstruction. By proposing the weighting strategy, we automatically perform 3D reconstruction to focus more on well-generated images, increasing the consistency of the 3D reconstruction.

We demonstrate a competitive performance of animatable 3D reconstruction compared to prior work on the Dynamic Furry Animals (DFA) datasets (Luo et al., 2022) and the GART dataset (Lei et al., 2024), despite using only a single image. We validate that DogRecon outperforms comparison methods on internet-sourced datasets in image-to-3D generation for dogs. In addition, DogRecon is seamlessly applicable to various applications, such as dog animations with pre-trained Gaussian scenes, and creating dog animation solely on text descriptions.

In summary, our contributions are as follows:

- We present DogRecon, a novel framework that aims to create an animatable 3D model of a dog from a single image, which is represented by a 3D Gaussian Mixture Model (GMM).
- We introduce a new approach that leverages a canine prior as guidance, enabling the generation of canine-centric multi-view images that are geometrically aligned with one another.
- We propose a new sampling strategy to mitigate consistency issues in zero-shot NVS. By weighting according to the alignment of the guidance mask with the generated images, we can create a more reliable 3D model. We validate that DogRecon outperforms comparison methods in image-to-3D generation for dogs. In addition, we demonstrate that DogRecon has comparable performance in animatable 3D reconstruction even if we use only a single image.
- We show that DogRecon has extensibility of two interesting tasks: animations with given scenes and text-to-animatable 3D reconstruction.

2 Related Work

In this section, we summarize related work on 3D reconstruction of animals, 3D generation from a single image, and animatable 3D reconstruction.

2.1 Animal 3D Reconstruction

Numerous attempts have been made to reconstruct 3D animals from 2D observation (Zuffi et al., 2019; Yang et al., 2022; Li et al., 2024; Luo et al., 2022; Sinha et al., 2023). Such approaches can be categorized into template-based methods and template-free methods. Template-based methods excel in leveraging predefined 3D shape and pose priors, facilitating accurate reconstructions even in ambiguous situations (Rüegg et al., 2023b, a). Specific parametric methods include SMAL (Zuffi et al., 2017) for quadrupeds, D-SMAL (Rüegg et al., 2023a) for dogs. Each parametric method can be used as a representation to infer the 3D shape

and pose from monocular images (Biggs et al., 2020; Rüegg et al., 2023b; Zuffi et al., 2019), and RGB-D images (Kearney et al., 2020). Recently, template-free approaches (Yang et al., 2022; Li et al., 2022; Yang et al., 2021; Yao et al., 2023; Li et al., 2024) offer broader applicability. While these methods are successful in obtaining models of various 3D shapes without specific templates, they are not explicitly editable to create novel motions.

In this work, we use template 3D shapes represented by D-SMAL (Rüegg et al., 2023a) as priors to optimize 3D Gaussians for creating an animatable dog model. By leveraging such geometric prior, we can explicitly control a 3D dog and generate novel dog pose animations (see Fig. 1).

2.2 Image-to-3D Generation

Advancements in 2D generative model (Rombach et al., 2022) and vision-language model (Radford et al., 2021) have pivoted towards assisting in the creation of 3D models (Poole et al., 2023). Following the success of the text-conditioned 3D model generation studies (Jain et al., 2022; Poole et al., 2023; Lin et al., 2023; Xu et al., 2023), there is a growing body of research on 3D model generation from a given image. In particular, studies that take a single image as input, Zero-1-to-3 (Liu et al., 2023b) predicts novel views from camera transformations and the input image, based on Stable Diffusion (Rombach et al., 2022). One-2-3-45 (Liu et al., 2023a) uses views generated by Zero-1-to-3 to obtain a 3D mesh by generalizable neural surface reconstruction. SyncDreamer (Liu et al., 2023c) generates multi-view consistent images leveraging synchronized multiview noise predictor. DreamGaussian (Tang et al., 2023) employs an efficiency-optimized framework, by leveraging generative 3D Gaussian to obtain a 3D model and using UV space enhancement to get a textured model. LRM (Hong et al., 2023) improves the generated image quality with a transformer-based architecture.

However, Zero-1-to-3 overlooks view consistency and a series of studies (Liu et al., 2023a; Tang et al., 2023) that aim to reconstruct complete 3D models by refining naive results from Zero-1-to-3, face significant limitations in maintaining view consistency. To resolve this issue, we propose a canine prior-guided NVS, which guides the generation process by silhouettes to synthesize geometrically aligned images (see Fig. 11).

2.3 Animatable 3D Reconstruction

Recent studies have made progress in 3D reconstruction by optimizing the canonical space of animatable objects using neural radiance fields (Weng et al., 2022; Peng et al., 2021b; Jiang et al., 2022; Huang et al., 2023) and 3D Gaussian (Lei et al., 2024; Hu et al., 2023a; Kocabas et al., 2023). These techniques have mainly focused on human models, with less work on animal models (Yang et al., 2022, 2023) and rely on leveraging multi-view images or videos to enhance detail. Recent efforts to data-efficiently reconstruct animatable 3D models have focused on single image input, which suffers from missing geometry and appearance due to partial observations. ELICIT (Huang et al., 2023) utilizes 3D body shape prior to guide optimization procedure. SHERF (Hu et al., 2023b) extracts hierarchical features of pixels and points, providing a strong clue to reconstruct the 3D model. Inevitably, addressed researches rely on already well-inferred SMPL parameters or captured 3D models corresponding to the given images in an efficient setup. In other words, they cannot be directly applied to the animal domain due to insufficient well-captured data. In addition, the 3D recovery performance of existing regressors for dogs (Rüegg et al., 2023a; Biggs et al., 2020; Rüegg et al., 2023b) is not guaranteed in difficult situations, such as a back view of a dog. The difficulty of the back view stems from the imbalance of the dataset and the inherent nature of dogs, where the dog's face is not visible in the back view.

In our work, we can reconstruct animatable 3D models of a dog by mitigating challenging cases of existing dog-specific regressors (see Fig. 3), enabling them to accurately find corresponding views across different views without the additional training data. We summarize a comparison of `DogRecon` with relevant animatable 3D reconstruction techniques in Table 1.

3 Method

In this work, we propose a new framework to reconstruct controllable 3D models of dogs from a single image, called `DogRecon`, that leverages canine geometry as prior (see Fig. 2). `DogRecon` comprises two main parts: Canine-centric NVS and 3D dog reconstruction with Reliable Sampling Weight. In the canine-centric NVS, we aim to generate multi-view images of an input image with mask guidance. This approach enables us to produce geometrically aligned images even for challenging cases, such as the back view of a dog (see Fig. 3). In the 3D dog reconstruction stage, we explicitly approximate a deformable Gaussian space from generated images of dogs. During the approximation process, we prevent using incorrectly generated images of dogs by employing adaptive sampling weights according to generated multi-view images. As a result, we can robustly optimize 3D dog Gaussian without blurry or cloudy effects (see Fig. 8).

Concretely, given a single 2D image x of a dog, we estimate a 3D dog template model M of a given dog image, represented by D-SMAL (Rüegg et al., 2023a). We generate multi-view images $\{\hat{x}_i\}$ by rotating and translating the image x using an image-conditioned diffusion model (Liu et al., 2023b). During the generation, we guide the diffu-

Table 1 Capacity comparison with relevant works. We divide the capabilities of prior relevant research into three factors to identify our proposed method. DogRecon is unique in that it can freely synthesize the 3D animation of the 3D model of a dog from just a single image, compared to other methods.

Method	Dog compatibility	Single image	Animatable
HumanNeRF (Weng et al., 2022)	×	×	✓
NeuMan (Jiang et al., 2022)	×	×	✓
BANMo (Yang et al., 2022)	✓	×	✓
LRM (Hong et al., 2023)	×	✓	×
DreamGaussian (Tang et al., 2023)	×	✓	×
GART (Lei et al., 2024)	✓	×	✓
ELICIT (Huang et al., 2023)	×	✓	✓
SHERF (Hu et al., 2023b)	×	✓	✓
DogRecon (ours)	✓	✓	✓

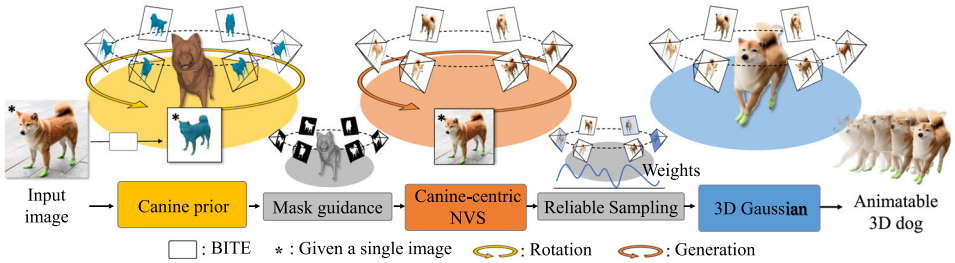

Fig. 2 Overview of DogRecon. Given a single dog image, we first predict the canine prior by BITE (Rüegg et al., 2023a). Based on the canine prior, we infer D-SMAL and the corresponding silhouette mask in the desired views, which guides the canine-centric NVS to generate canine-centric multi-view images. Finally, we create an animatable 3D Gaussian dog with a Reliable Sampling Weight.

sion process according to a silhouette mask $\hat{m}_i$ obtained by explicitly rotating the 3D model of the dog M. We then optimize the deformable Gaussian Mixture Model (GMM) using the given image with predicted D-SMAL pair, and generated images with rotated and translated D-SMAL pairs $\{(x, M), (\hat{x}_1, \hat{M}_1), ..., (\hat{x}_N, \hat{M}_N)\}$. To robustly reconstruct the 3D Gaussian dog, we utilize a sampling weight W that can handle poorly generated images. After the reconstruction, we can explicitly control the 3D Gaussian dog because we utilize a template 3D model of a dog and also generate novel motions with texture.

3.1 Preliminary

D-SMAL. D-SMAL (Rüegg et al., 2023a) is a dog-specific version of a parameterized 3D model derived from SMAL (Zuffi et al., 2017), which depends on the shape and pose of the animal. D-SMAL uses a template dog mesh in a neutral pose to deform the vertices into a space formed within the template coordinate system. D-SMAL consists of the dog-specific shape parameters $\beta \in \mathbb{R}^{30}$ to define the body shape of a dog. D-SMAL also includes the pose parameters $\theta \in \mathbb{R}^{35\times3\times3}$ to specify the angles of joints according to the tree-like hierarchy structure, and the global position parameter $\gamma \in \mathbb{R}^3$ to place the dog model in space. Here, $\theta_0 \in SO(3)$ indicates the root pose, while the remaining pose parameters represent the rotation of each body joint relative to its parent.

There are two benefits of adapting D-SMAL in our framework. Firstly, we use D-SMAL as the explicit canine representation to generate an animatable 3D dog for various dog breeds. Secondly, D-SMAL allows us to give guidance to generate canine-centric multi-view images. For this reason, we leverage D-SMAL in our framework.

3.2 Canine-centric NVS with Canine Prior

In this section, we focus on generating multi-view images guided by the silhouette mask $\hat{m}$ obtained from an explicitly rotated 3D model to get geometrically aligned images. We set the D-SMAL parameter and its silhouette mask in the desired view by utilizing a 3D model of the input image, called a canine prior.

Before addressing the canine prior, it is worth noting that we cannot directly apply the 3D dog pose estimator to all views. The lack of sufficient training data and the structural complexity of dogs often lead to failures in certain views, particularly in handling back views (see Fig. 3). If we fail to predict the 3D dog pose in certain views, we cannot use those views for reconstruction, as it will result in a blurred effect. To handle this issue, we propose an explicit approach

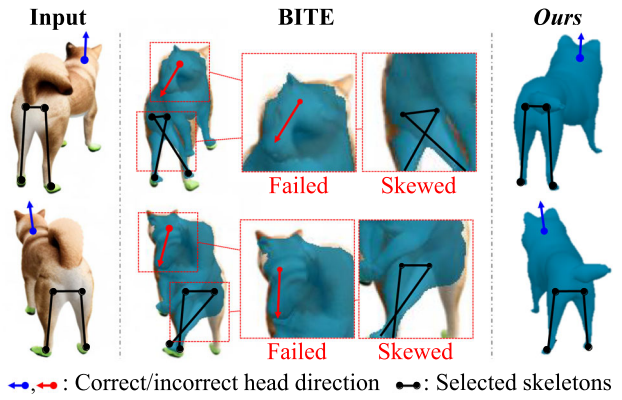

Fig. 3 Failure cases of pose regressor. We regress 3D models of dogs from given images with off-the-shelf dog pose regressor (Rüegg et al., 2023a). As shown in each red zoom-in box, the face of the 3D model is oriented in the wrong direction, and the legs may skew. Using such incorrect 3D models to optimize Gaussian space can result in significantly noisy textures

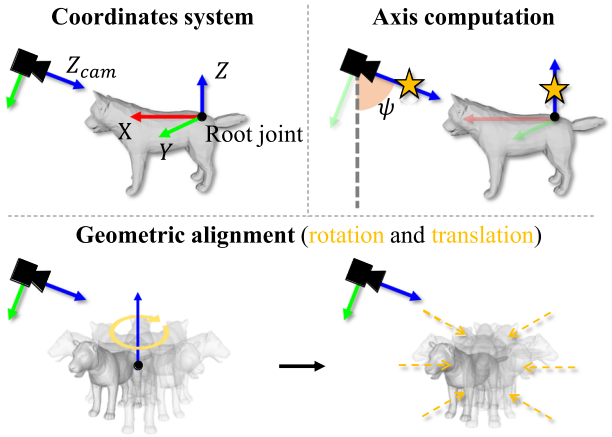

Fig. 4 Illustration of explicit rotation with canine prior. *Top left:* Coordinates systems of dog and camera. *Top right:* We determine the rotation axis using the angle ψ. *Bottom:* Then we rotate the model M to settle the D-SMAL at the desired view and translate it. We render this D-SMAL mesh to the desired view and use it as a prior for NVS.

to obtain the correct 3D model for the generated image of a dog.

Canine Prior. Given a single dog image $x \in \mathbb{R}^{H\times W\times 3}$, we first estimate D-SMAL to extract canine prior information by BITE (Rüegg et al., 2023a). In this work, we call initial D-SMAL as *canine prior*, which is obtained from an input image x. Here, we assume that we can estimate D-SMAL reliably, at least for the input image. Given a canine prior estimated from the input image x, the 0-th pose θ_0 in D-SMAL represents the root pose of the dog, located at the pelvis. As illustrated in the top left of Fig. 4, the coordinate system of a dog is defined with X, Y, and Z axes centered at this root joint. Assuming the dog is on the ground with the canonical pose, we can define a canine-centric rotation axis by computing the angle ψ between the z-axis of the dog coordinate and the principal axis of the camera coordinate. Then, we can infer the D-SMAL parameter for the desired viewing angle by rotating θ_0 based on the canine-centric rotation axis. Here, it should be noted that the origin of the root pose is located at the pelvis, not the center of mass. Thus, it is difficult to estimate a canine-centric D-SMAL for canine-centric NVS directly. For example, a misaligned rotation axis causes a dog to appear to be sliding or floating around the pivot point rather than turning in place at the desired view. To correct this misalignment, we compute a 3D translation between the pelvis and the center of keypoints composing D-SMAL and reflect this translation into the global location γ of D-SMAL (see bottom of Fig. 4). Finally, the adjusted canine-centric D-SMAL $\hat{M}$ is used to render the corresponding silhouette mask $\hat{m}$ in the 2D image domain. By doing so, we can leverage the silhouette mask $\hat{m}$ as guidance for NVS and settle D-SMAL for generated images, one of our key contributions.

Canine-centric NVS. We choose Zero-1-to-3 (Liu et al., 2023b) as the backbone network for our canine-centric NVS. Zero-1-to-3 can generate a synthesized image according to the desired viewpoint, suitable for aligning with the explicitly rendered silhouette $\hat{m}$. Zero-1-to-3 takes a single image of an object and the desired camera pose as input and then synthesizes a new image for the desired viewpoint:

$$\hat{x}_{R,T} = f(x, R, T), \tag{1}$$

where $f(\cdot)$ is the pre-trained Zero-1-to-3, $\hat{x}_{R,T}$ is a synthesized image for the relative rotation R and translation T of the desired viewpoint. Zero-1-to-3 relies on a conditional diffusion model (Rombach et al., 2022) and specifically uses the embedding of the input view and relative camera extrinsic as conditions to iteratively denoise the Gaussian noise on latent space:

$$\tilde{z}_{t-1} = \epsilon(z_t, t, c(x, R, T)), \tag{2}$$

where $\epsilon(\cdot)$ is an U-Net denoiser, z_t is the latent at the diffusion time step t and $c(x, R, T)$ denotes a CLIP embedding (Radford et al., 2021) for x, R and T. Here, we denote a denoised latent as $\tilde{z}_{t-1}$, which is tentative latent for mask guidance. However, as addressed in prior work (Ye et al., 2023), Zero-1-to-3 is not guaranteed to generate geometrically aligned multi-view images. This could be seen as a result of insufficient training data that captures the variability of deformable objects.

To address this issue, we propose a canine-centric NVS module, which leverages the estimated D-SMAL and the corresponding silhouette mask $\hat{m}$ as guidance for NVS (see Fig. 5).

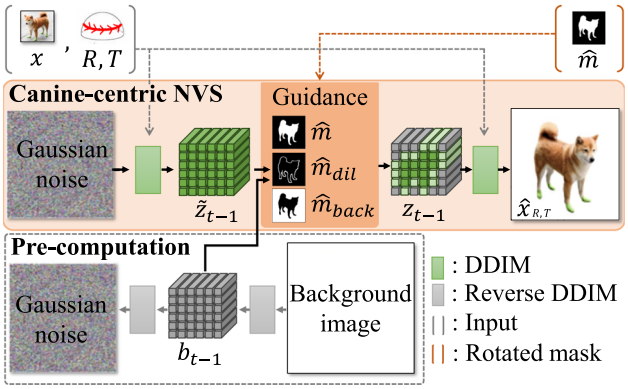

Fig. 5 Canine-centric NVS module. The input consists of x, R, T and $\hat{m}$, while the output is $\hat{x}_{R,T}$. In the guidance part (orange box), we leverage the white region of each mask $\hat{m}$, $\hat{m}_{dil}$ and $\hat{m}_{back}$ to fuse the two latent vectors $\tilde{z}_{t-1}$ and b_{t-1} (Color figure online).

Inspired by DIFFEDIT (Couairon et al., 2023), we interpolate region-specific feature vectors in the latent space during the denoising step. Each region is distinctly separated into three parts: the original mask $\hat{m}$, the dilated mask $\hat{m}_{dil} \in \mathbb{R}^{H/8\times W/8}$ (acting as a buffer zone), and the background $\hat{m}_{back} \in \mathbb{R}^{H/8\times W/8}$, ensuring there is no overlap between them (see Fig. 5). We dilate the silhouette mask $\hat{m}$ to make $\hat{m}_{dil}$ and $\hat{m}_{back}$. Before the denoising steps, we pre-compute a background latent b by leveraging reverse denoising steps to a pure white image. Our canine-centric NVS fuses the two latent vectors, $\tilde{z}_{t-1} \in \mathbb{R}^{H/8\times W/8\times C}$ for the dog and $b_{t-1} \in \mathbb{R}^{H/8\times W/8\times C}$ for the background image, in three separate mask regions:

$$z_{t-1} = \begin{cases} \tilde{z}_{t-1} & \text{if } u, v \in \hat{m}, \\ \alpha\tilde{z}_{t-1} + (1-\alpha)b_{t-1} & \text{if } u, v \in \hat{m}_{dil}, \\ b_{t-1} & \text{if } u, v \in \hat{m}_{back}, \end{cases} \tag{3}$$

where α is weight factor for fusing $\tilde{z}_{t-1}$ and b_{t-1}. We gradually reduce the dilated mask region and weight factor α during denoising to guide the image to fit the silhouette mask. Leveraging $\hat{m}_{dil}$ with α helps generate a more geometrically accurate image of the dog. With this mask guide process, the proposed canine-centric NVS $f_{canine}(\cdot)$ generates a canine-centric image for desired viewing angle:

$$\hat{x}_{R,T} = f_{canine}(x, R, T, \hat{m}, b). \tag{4}$$

Using Eq. (4), we generate the canine-centric images that cover a 360-degree view of the canine. It should be worth noticing that without fine-tuning to any dataset, our canine-centric NVS can mitigate the limitation of the original NVS (Liu et al., 2023b), which struggles to generate geometrically accurate views for real dogs. In addition, our NVS model, which directly rotates the 3D model to guide generation, can be extended to various quadrupeds and humans. Detailed alignment of D-SMAL for reconstruction as described in Sec. 6.2.

3.3 3D Reconstruction with Sampling

In this section, we address the approximation of 3D Gaussians from multi-view image paired with 3D template models $\{(x, M), (\hat{x}_1, \hat{M}_1), ..., (\hat{x}_N, \hat{M}_N)\}$. To reconstruct a 3D dog robustly, we propose a weighting strategy that concentrates on reliably generated images.

We leverage the fundamental concepts of the Gaussian Articulated Template Model (GART) (Lei et al., 2024) to represent and reconstruct articulated subjects, specifically focusing on dogs to create a 3D animatable dog. At the core of this representation is the Gaussian Mixture Model (GMM), which effectively approximates the 3D shape and texture of a dog. Each Gaussian in the mixture is defined by parameters such as its mean (position in 3D space), covariance matrix (which controls its size, shape, and orientation), and a color component representing the appearance of the corresponding part of the model. These parameters are optimized to match the observed data, capturing the articulated movement of dogs through a predefined skeletal structure informed by D-SMAL. We can express these Gaussian parameters as the output of GMM that takes pose θ of D-SMAL as input.

The deformation of the model to match observed frames is achieved through a process of differentiable rendering, known as 3D Gaussian Splatting (Kerbl et al., 2023). This technique projects the 3D Gaussian onto the image plane, allowing for the gradient-based optimization of the Gaussian parameters against the observed images. This process ensures that the reconstructed model accurately reflects the appearance and motion of the subject in the input, providing a highly detailed and dynamic representation.

Reliable Sampling Weight. Conventional regression loss functions in Neural Radiance Field (Mildenhall et al., 2020) and 3D Gaussian Splatting (Kerbl et al., 2023) assume a training dataset comprised entirely of real images. However, our scenario involves only a single real image, with the rest of the training data being generated images. This difference makes the simple regression loss unsuitable for achieving consistent 3D reconstruction. To address the challenge of optimizing our model when the training dataset comprises primarily generated images, with only a single real image as a reference, we introduce the concept of Reliable Sampling Weight (RSW). RSW adjusts the influence of each sample during the optimization process, giving higher priority to samples that exhibit geometric and semantic consistency according to the reference image. We derive the weights for each generated image by measuring how similarly it aligns with the real image, using CLIP-Similarity (Jain et al., 2021) and the

L1 difference of silhouette masks:

$$W = 1 + \lambda_{CLIP}\phi(x)^T\phi(\hat{x}) - \mathcal{L}_1(m', \hat{m}), \tag{5}$$

where W is the RSW that applies to each training sample (i.e., one scalar value per sample), x is a real input image, $\hat{x}$ and $\hat{m}$ are the generated image and silhouette mask by the proposed canine-centric NVS, and m' is the pseudo mask for $\hat{x}$ predicted by Segment Anything Model (SAM) (Kirillov et al., 2023). The term $\lambda_{CLIP}\phi(x)^T\phi(\hat{x})$ captures the cosine similarity between the real and generated images using the normalized embeddings of CLIP-VIT (Radford et al., 2021) (we set λ_{CLIP}=1). The term $\mathcal{L}_1(m', \hat{m})$ measures the discrepancy between the predicted mask m' and the rendered mask $\hat{m}$ from canine prior. A lower value of $\mathcal{L}_1$ indicates plausible geometry. The silhouette L1 term is normalized to ensure $\mathcal{L}_1(m', \hat{m}) \in [0, 1]$, guaranteeing that $W > 0$. When silhouette errors are large, W decreases the contribution of the RGB loss, preventing overfitting to incorrect geometries. Based on the proposed RSW, the overall training loss is defined as:

$$\mathcal{L} = W\mathcal{L}_1(\tilde{x}, \hat{x}) + \lambda_{SSIM}\mathcal{L}_{SSIM}(\tilde{x}, \hat{x}) + \mathcal{L}_{reg}, \tag{6}$$

where $\tilde{x}$ denotes the rendered image from 3D Gaussians, $\mathcal{L}_{SSIM}$ and λ_{SSIM} are the SSIM loss and weight, and $\mathcal{L}_{reg}$ is regularization term, similar to GART (Lei et al., 2024).

RSW reduces the influence of anomalous samples with incorrect geometries or visual artifacts by selectively weighting them according to reliable semantic and geometric properties. This mechanism ensures that training focuses on consistent and plausible data, enhancing the robustness and fidelity of the reconstructed 3D representations. As shown in Fig. 8, RSW effectively balances semantic and geometric consistency, enabling reliable 3D reconstructions from a single real image.

4 Experiments

In this section, we mainly address the evaluation and comparison of `DogRecon` over comparable methods. We provide information for the experimental setup (Sec. 4.1) and describe details of comparison methods (Sec. 4.2). Our experiments can be divided into image-to-3D generation and animatable 3D reconstruction tasks (Sec. 4.3). Then, we address the ablation study (Sec. 4.4) and show the applicability of the proposed `DogRecon` on two applications (Sec. 4.5).

4.1 Experimental Setup

Implementation details. DogRecon constructs an animatable 3D dog using a Gaussian Mixture Model (GMM). The Gaussian dog reconstruction stage takes 10K iterations for optimization. This training phase takes approximately 6 minutes on only one NVIDIA RTX 4090 GPU. We follow the parameter settings of GART (Lei et al., 2024).

Dataset. To evaluate the animatable 3D reconstruction task, we utilize the GART dataset (Lei et al., 2024), the Dynamic Furry Animals (DFA) dataset (Luo et al., 2022).

GART (Lei et al., 2024) proposes a dataset for experimenting with 3D reconstruction of dogs. The GART dataset consists of in-the-wild monocular videos capturing eight dog breeds. The videos are divided into a training set featuring dynamic movements and a test set with relatively less movement. We select four breeds to experiment with, excluding those with similar shapes and textures.

ARTEMIS (Luo et al., 2022) introduces the DFA dataset for experiments in 3D animal reconstruction. The DFA dataset contains multi-view renderings and skeletal motions of nine high-quality computer-generated (CGI) furry animals. In our experimental setup, we use different cameras for the training and test sets. We select two quadrupeds, the wolf and the beagle, as our subjects for experimentation.

To experiment the image-to-3D generation task, we only utilize the internet-sourced dog dataset (Dognomics2023, 2023) for comparison. The internet-sourced dog dataset consists of a single RGBA image for each breed of real dogs.

Evaluation metrics. To evaluate 3D generation performance, we adopt the CLIP cosine similarity metric (Radford et al., 2021). In our experimental setup, we utilize real dog images as input, which means that ground truth 3D models of the dogs are not available. Therefore, following prior work (Tang et al., 2023), we use the CLIP cosine similarity between the input image and the target view rendered from the obtained 3D models. To compare the quality of reconstructed 3D models, we mainly use PSNR, SSIM, and LPIPS (Zhang et al., 2018), as metrics to capture different aspects of photo-realistic similarity.

4.2 Baselines

We implement and modify the prior works to compare our `DogRecon` with them. For image-to-3D generation, we compare our method with conventional image-conditioned 3D generation approaches. For animatable 3D dog reconstruction, no exact methods use a single dog image as input. Therefore, we compare `DogRecon` with the video-based methods, GART (Lei et al., 2024) and BANMo (Yang et al., 2022). Detailed implementation is provided in the following paragraphs.

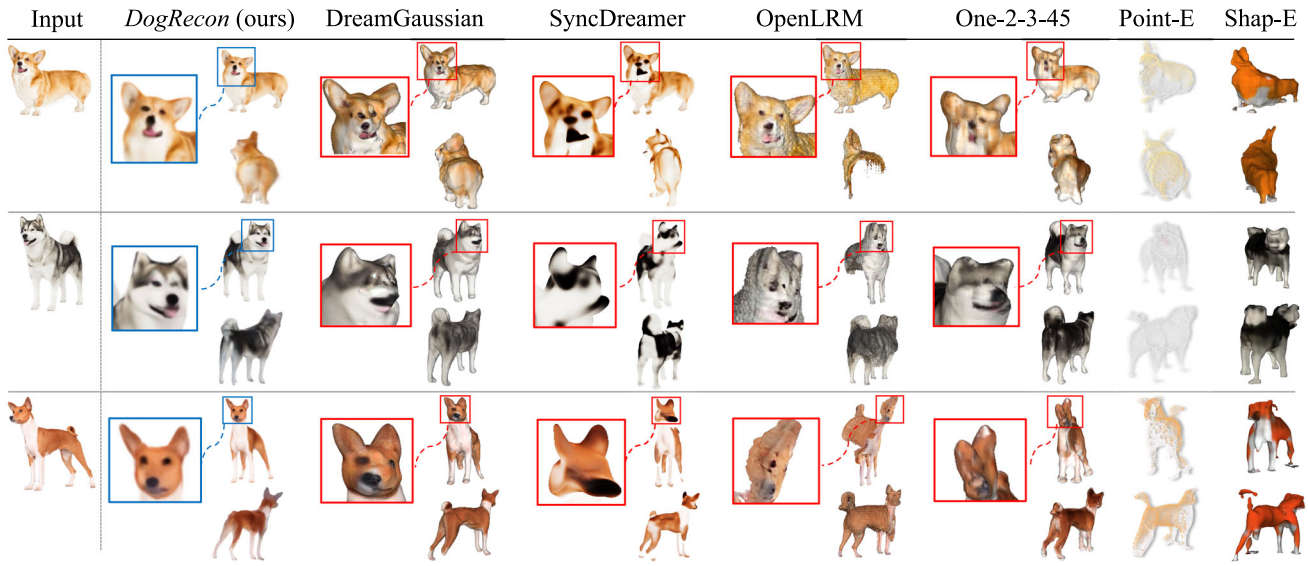

Fig. 6 Qualitative comparison of image-to-3D generation. We compare the rendering quality on the internet-sourced dog dataset. Blue and red boxes are zoomed-in views that show more detailed textures. The comparison highlights the differences in texture quality and structural consistency in the generated 3D shapes. Our method outperforms others in maintaining consistent textures and finer details, especially in close-up regions (Color figure online).

Table 2 Quantitative comparison of image-to-3D generation on Internet-sourced dataset. For DreamGaussian (Tang et al., 2023), a mesh fine-tuning stage is used to improve quality. `DogRecon` outperforms other methods

Method	CLIP-Similarity ↑
Point-E (Nichol et al., 2022)	0.6789
Shap-E (Jun & Nichol, 2023)	0.7279
One-2-3-45 (Liu et al., 2023a)	0.7520
OpenLRM (Hong et al., 2023; He & Wang, 2023)	0.7731
SyncDreamer (Liu et al., 2023c)	0.7783
DreamGaussian (Tang et al., 2023)	0.7905
`DogRecon` (ours)	**0.8006**

GART (Lei et al., 2024). GART (Gaussian Articulated Template Model) is an explicit representation for capturing and rendering non-rigid articulated objects from monocular videos, utilizing 3D Gaussians to approximate geometry and appearance. We train GART for each dog subject by optimizing the model on the training data of more than 170 frames. However, GART uses a lot of frames, so a fair comparison is challenging. Therefore, we implement Few-shot GART†, using only 3 distinct frames that capture various poses of the dog, ensuring diversity in the input data. We primarily compare the performance of our method with Few-shot GART† under these conditions.

BANMo (Yang et al., 2022). BANMo reconstructs implicit 3D models from multiple casual RGB videos. We only compare our method with BANMo without a few-shot version of BANMo. There are two reasons. Firstly, BANMo cannot be trained with 3 frames. This is because it usually implicitly learns deformable 3D models, requiring more than 800 images for realistic reconstruction performance in general. Secondly, BANMo uses the optical flow between frames for training, which is difficult to obtain with 3 frames. Thus, we do not compare to BANMo with a few shot settings, but rather train only the naive BANMo on our dataset, as we believe it is sufficient to compare to the existing naive BANMo.

4.3 Comparison

In this section, we evaluate two distinct tasks. First, image-to-3D generation aims to reconstruct a static 3D model from a only single 2D image. Second, animatable 3D reconstruction not only reconstructs the 3D shape but also infers a skeletal structure, enabling novel pose rendering. We use the internet-sourced dog dataset to assess image-to-3D generation in in-the-wild scenarios. For animatable 3D reconstruction, we employ the GART (Lei et al., 2024) dataset and the DFA (Luo et al., 2022) dataset, both of which contain consecutive video

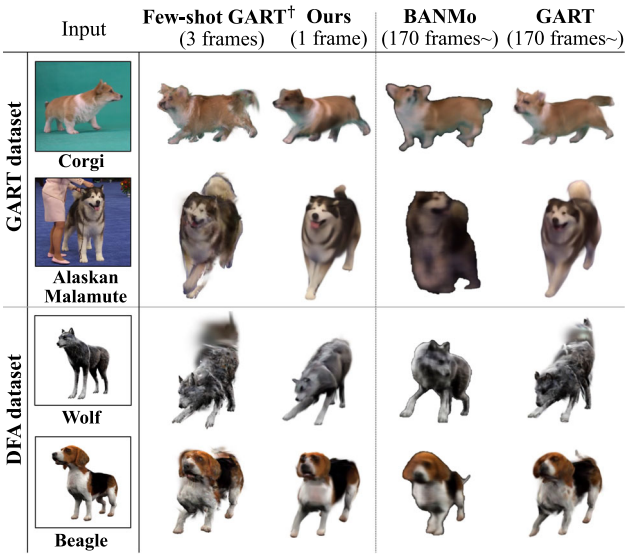

Fig. 7 Qualitative comparisons of animatable 3D reconstruction. We qualitatively compare the rendering quality of novel poses on the GART dataset and the DFA dataset

frames or sufficient multi-view data for testing novel pose and novel view synthesis.

Image-to-3D generation. We qualitatively compare DogRecon with state-of-the-art methods on the internet-sourced dog dataset (see Fig. 6). The internet-sourced dog dataset consists of in-the-wild images, with out-of-distribution examples not present in the training data, making it a challenging setting. Point-E (Nichol et al., 2022) and Shap-E (Jun & Nichol, 2023) have poor generation quality for in-the-wild dog images. OpenLRM (He & Wang, 2023; Hong et al., 2023) and One-2-3-45 (Liu et al., 2023a) suffer from low-grade generation quality, resulting in distorted or blurred facial features. SyncDreamer (Liu et al., 2023c) is losing texture clarity and causing facial structural artifacts. DreamGaussian (Tang et al., 2023) shows unnatural distortions in the head and face structure. In contrast, our DogRecon outperforms in terms of color, geometric structure, and texture quality compared to other methods. DogRecon preserves fine texture details, especially in close-ups, while other methods struggle to maintain accuracy in features like eyes.

Table 3 Quantitative comparisons of animatable 3D reconstruction on the GART dataset

	Alaskan Malamute			Corgi		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GART (Lei et al., 2024)	17.84	0.813	0.242	23.41	0.928	0.093
Few-shot GART†	15.51	0.736	0.323	**20.04**	0.880	**0.125**
DogRecon (ours)	**16.86**	**0.818**	**0.228**	19.84	**0.901**	0.126
	German Shepherd			French Bulldog		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GART (Lei et al., 2024)	16.62	0.812	0.233	19.18	0.849	0.229
Few-shot GART†	14.38	0.748	0.274	17.72	0.803	0.242
DogRecon (ours)	**15.87**	**0.798**	**0.258**	**19.07**	**0.845**	**0.237**

Table 4 Quantitative comparisons of animatable 3D reconstruction on the DFA dataset

	Wolf			Beagle		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GART (Lei et al., 2024)	16.70	0.796	0.177	15.45	0.797	0.224
Fewshot-GART†	14.94	0.769	**0.195**	14.37	0.778	0.228
DogRecon (ours)	**16.53**	**0.811**	0.196	**23.06**	**0.913**	**0.102**

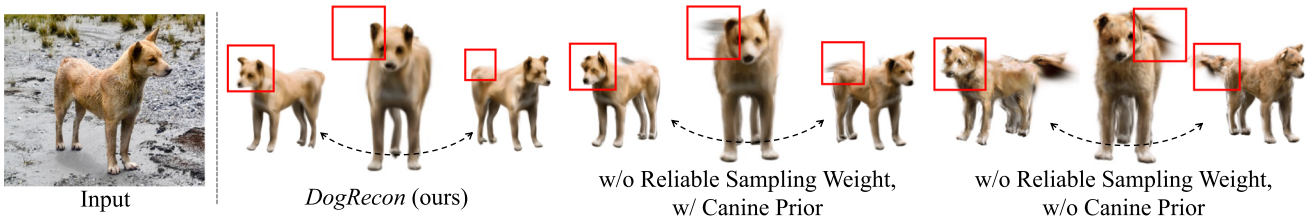

Fig. 8 Ablation study on each module. We show the rendered images of our model from an input image. We can observe a smooth surface in the case of using the canine prior (different between the third and fourth columns) and more complete geometry in the case of using Reliable Sampling Weight (different between the second and third columns)

Fig. 9 More results of DogRecon. We show the results of DogRecon on images of various breeds of dogs. For more detailed results, check out the supplementary material video.

For the quantitative comparisons, we measure the CLIP-similarity (Radford et al., 2021) between the input image and the images rendered from the generated 3D models in novel views (see Table 2). Our DogRecon shows the highest CLIP-similarity results compared to other baselines for in-the-wild dog images. Based on DietNeRF (Jain et al., 2021), views of the same semantic object exhibit high cosine similarity. Consequently, our DogRecon produces the most semantically consistent results.

Animatable 3D reconstruction. We qualitatively compare DogRecon with state-of-the-art methods on the GART and DFA datasets. We show the rendering quality of each model to the novel pose based on the input image (see Fig. 7). To compare with BANMo (Yang et al., 2022), we selected frames from the input sequence that closely matched the target novel pose. BANMo struggles to manipulate the reconstructed 3D model into the desired pose. GART (Lei et al., 2024) utilize over 170 input frames, so textures and details are well reproduced. In contrast, Few-shot GART†, using only 3 frames, achieves reasonable quality but lacks detail. Our DogRecon achieves better quality with just a single frame. For example, with the Alaskan Malamute and Beagle subjects, Few-shot GART† generates somewhat blurry textures and lacks fine details, whereas DogRecon provides clearer, more detailed textures with just one frame. Furthermore, DogRecon demonstrates superior structural accuracy, particularly in regions like the face and legs, and renders poses more naturally compared to Few-shot GART†.

For the quantitative comparisons, we evaluate selected four subjects from the GART dataset and two from the DFA dataset, as shown in Table 3 and Table 4, respectively. Overall, DogRecon demonstrates better performance than Few-shot GART†. On the GART dataset, DogRecon shows comparable performance compared to GART. DogRecon trained on only one real frame performs similarly to GART trained on over 170 frames. In specific cases, such as the Corgi subject, the performance difference can be attributed to differences in input conditions, particularly the fact that the Corgi's tail is not visible in the input image (see Fig. 7). This leads to suboptimal results for DogRecon in this instance. On the DFA dataset, we show that DogRecon demonstrates better performance than GART when the view distribution of the training set and the view distribution of the test set are not similar, *i.e.,* extrapolation Novel View Synthesis. The performance difference between DogRecon and GART arises from discrepancies in camera location and viewing angle, as training and testing use data captured from different cameras. The dog's pose remains consistent during training and testing, so the performance difference is due to viewpoint variations rather than pose discrepancies. We show more results of DogRecon in Fig. 9.

Fig. 10 Applications of DogRecon. (a) Novel animation within pre-trained Gaussian scene. (b) Novel pose animation from a given text prompt.

4.4 Ablation Study

Ablation study on each module. As shown in Fig. 8, we show that DogRecon with canine prior and RSW (second column) is better preservation of geometry compared to the partial model. The proposed RSW module can assist the optimization process by adaptive weight of each viewpoint which maintains consistency (different between the second and third columns). The proposed model with the canine prior is a better texture representation than the naive model without any bumpy surface and fog-like artifact (different between the third and fourth columns). From these results, we believe that each module effectively synthesizes the complete and realistic 3D geometry of dogs.

Effectiveness of mask guidance. As shown in Fig. 11, we introduce canine guidance to refine geometric alignment in our canine-centric NVS module by overlapping the naive NVS results with an explicitly rotated canine prior. This approach can occasionally omit slender details (e.g., portions of the dog's tail), largely due to the skeletal limitations of D-SMAL, which is optimized for rigid alignment rather than highly deformable features. Nonetheless, canine guidance substantially improves the reconstruction of broader canine structures such as the legs, head, and torso, reflecting a notable trade-off between overall alignment consistency and fine-grained tail depiction. As shown in Fig. 11, blue squares show having a binary mask without a dilated mask sometimes results in images that follow canine guidance. However, the absence of α ("w/o α") results in outputs similar to the naive NVS, which lack proper alignment with canine guidance. This highlights the crucial role of α in refining geometric alignment, as it ensures smoother transitions and mitigates artifacts when blending features during the denoising process (see red circles in Fig. 11. Our canine-centric NVS achieves improved geometric consistency with the dilated mask ("with α"), highlighting the importance of this parameter for accurate and trustworthy reconstructions.

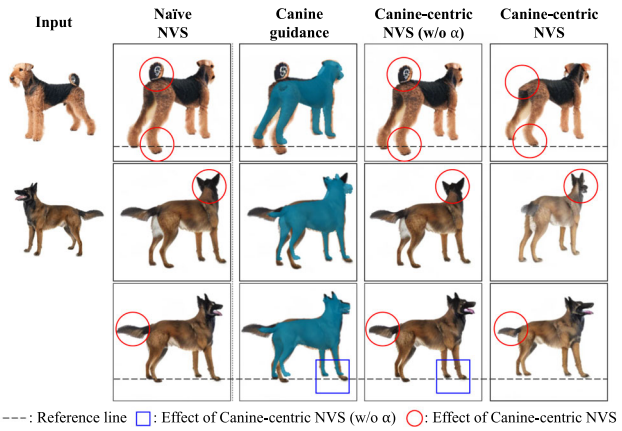

Fig. 11 Effectiveness of canine-centric NVS. *Naive NVS:* Generated results from Zero-1-to-3 (Liu et al., 2023b). *Canine guidance:* Overlap between naive result and explicitly rotated canine prior. *Canine-centric NVS (w/o α)*: Generated results from the proposed canine-centric NVS without the parameter α. *Canine-centric NVS:* Generated results from the proposed canine-centric NVS. Blue squares show the effect of the binary mask of canine-centric NVS (without dilated mask). Red circles highlight improvements in the canine-centric NVS (with α) (Color figure online).

4.5 Application

We show the applicability of DogRecon for two interesting applications: Animatable 3D dog within scene and text-to-animation (see Fig. 10). Application details that realistically create various pose animations and composite them into pre-trained Gaussian scenes are available in the supplementary material video.

Text-to-Animation. As DogRecon uses only a single image, we can generate previously unseen images by utilizing

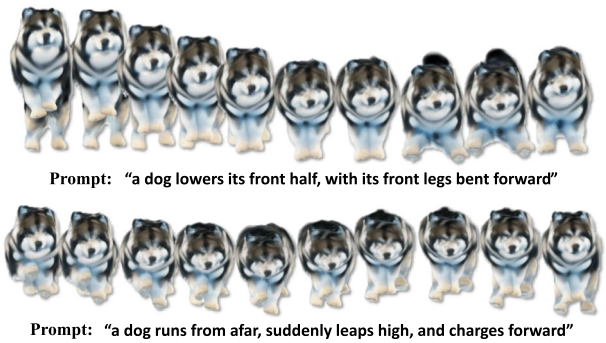

Fig. 12 Application of DogRecon: Text to motion. We generate a motion with the following prompt to OmnimotionGPT (Yang et al., 2024) and then animate with a 3D Gaussian dog with DogRecon.

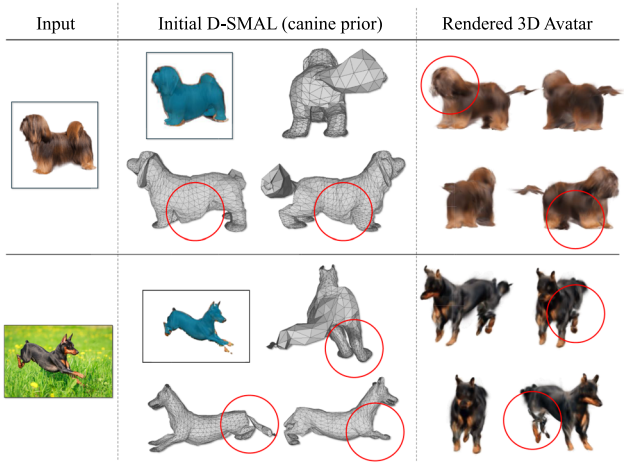

Fig. 13 Failure cases. The first column shows each single image input. The second column displays the initial 3D shapes with invalid predictions (highlighted in red circles). The third column presents the rendered novel pose and views of 3D avatars corresponding to each input image (unnatural effects are highlighted in red circles) (Color figure online).

a text-conditioned image generation model such as DALL.E-2 (Ramesh et al., 2022). We infer a 3D model from the generated images (see Fig. 10). Furthermore, by using text-to-motion generation method called OmnimotionGPT (Yang et al., 2024), we can generate the motion of the obtained 3D model (see Fig. 12). This approach allows for editing explicit animation of the 3D models, showing broader applicability.

Animatable 3D dog within scene. DogRecon can synthesize complete and realistic novel pose animation in pre-trained Gaussian scenes. We directly placed the DogRecon model within the scene, allowing us to generate animations by manipulating the pose and joint parameters of the 3D dog model. The results show that our framework can facilitate combining Gaussian scenes without any additional training.

4.6 Discussion

In our experiments, we observe that most failure cases appear when it is difficult to estimate the pose or shape of the input image. As shown in the first row of Fig. 13, failing to capture the accurate shape from the input image hinders learning finer texture and instead produces bulky or blurry rendered images. Similarly, as shown in the bottom row of Fig. 13, highly dynamic and uncertain poses make it difficult to learn local textures, such as those on the paws.

Our framework allows us to reconstruct a 3D dog avatar from a single image. Nevertheless, our framework can still fail under certain conditions. First, our framework may not reconstruct a dog's tail when the tail is complex or partially visible. Additionally, our framework produces a blurry 3D Gaussian dog if the input dog has highly dynamic or uncertain poses as shown in the bottom row of Fig. 13. Overcoming these limitations would be an exciting avenue for future work.

5 Conclusion

We have presented DogRecon, a novel framework that can reconstruct an animatable 3D dog Gaussian from a single dog image. We newly introduce canine-centric NVS with canine prior, which allows us to generate a set of canine-centric multi-view images that fits the geometry of dogs. Furthermore, 3D reconstruction with the proposed RSW enables to complement of a certain level of uncertainty in the canine-centric NVS. Through the synergy of the two proposed modules, DogRecon can generate animatable 3D dogs from a single image. In addition, we show the applicability of DogRecon via two interesting applications. As a future direction, animatable 3D reconstruction of non-canine animals beyond dogs would be fascinating.

6 Appendix

6.1 Consistency of DogRecon

To evaluate the consistency of proposed DogRecon, we show and assess the generated 3D model with the input reference image, both qualitatively and quantitatively. As shown in Fig. 14, we render 19 images with each 3D dog avatar rotating clockwise in 18-degree increments, starting with the left-facing image. We report the CLIP cosine similarity to measure the consistency between the reference image and the respective rendered image. Fig. 14 shows high and consistent CLIP similarity scores, which indicate consistency between the input and the rendered 3D dog avatars. That is,

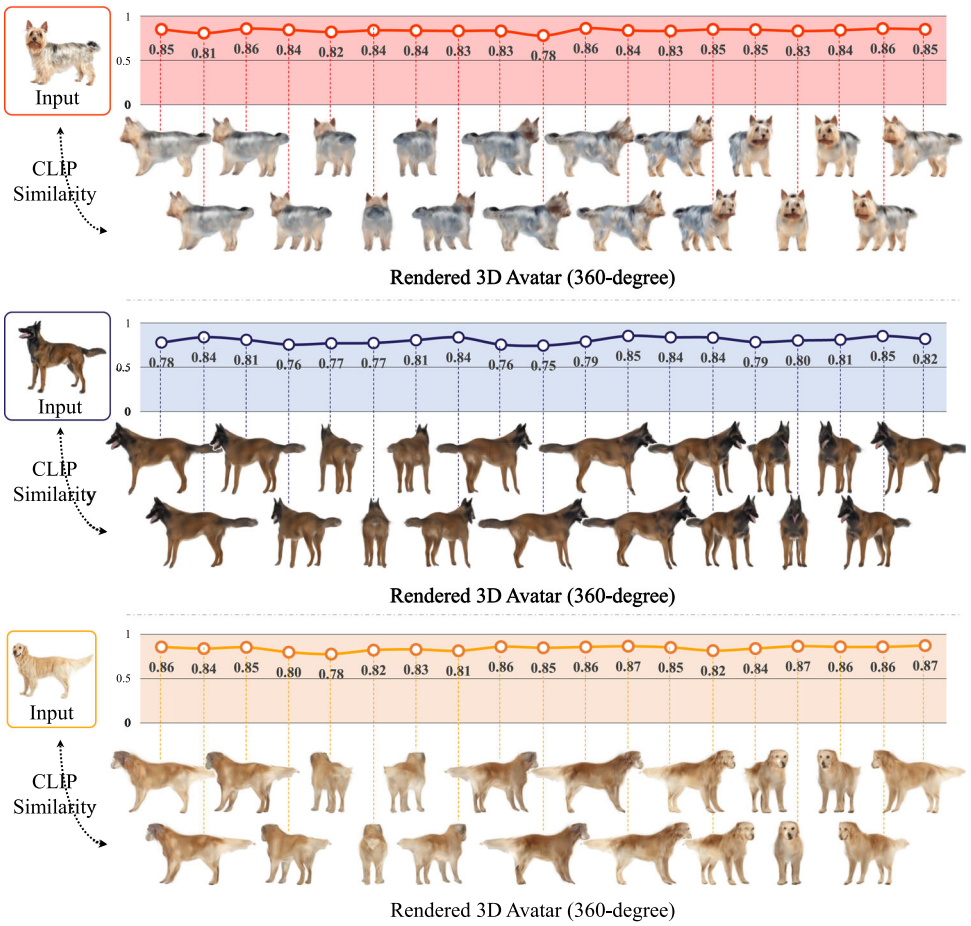

Fig. 14 Qualitative and quantitative results on rendered 3D dog avatars. For three input dog images (left), 360-degree views of their rendered 3D dog avatars are shown with corresponding CLIP similarity scores (y-axis) across viewpoints (x-axis).

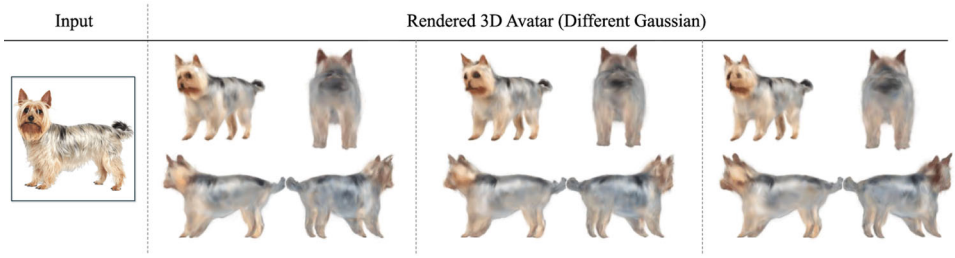

Fig. 15 Diversity in rendered 3D dog avatars. Qualitative results showing the diversity of rendered 3D avatars based on different random seeds. The input image is shown on the left, followed by three columns of rendered 3D avatars generated from differently optimized 3D Gaussians.

the generated avatar aligns consistently with the reference image.

We evaluate the diversity of the generated 3D dog model from the same reference image using different seeds. Our framework is intentionally designed to produce consistent results when provided with the same input image. A fixed seed is used to create the results of all figures except Fig. 15. However, to show the diversity of our framework, we include results with varying seeds in Fig. 15. Even when using different seeds, the rendered images of the 3D Gaussian dog remain consistent with the reference image.

6.2 D-SMAL alignment

Parametric models typically rely on highly accurate parameters obtained from controlled motion capture systems for 3D reconstruction. However, this level of precision is not achievable in the wild environments. In such scenarios, we perform per-image alignment of the parametric model for each training instance, as seen in recent studies (Jiang et al., 2022; Lei et al., 2024). Our canine-centric NVS module handles the geometric reconstruction, but the initial D-SMAL parameters for the canine prior are often inaccurate. This is primarily because zero-shot NVS models struggle to produce consistent outputs across diverse, unconstrained inputs. Despite the geometric reconstruction guided by the Canine-centric NVS,

inconsistencies in the D-SMAL canine prior lead to suboptimal alignment. To address these challenges, we refine the D-SMAL parameters for each training image by aligning the model outputs with the observed input data. Given an image, we obtain the segmentation mask m' of the dog using the Segment Anything Model (SAM) (Kirillov et al., 2023). We further set the D-SMAL parameters $\hat{M}$ of generated images, a collection of vertices and faces, based on rotating and translating the canine prior. Since the initial estimates of θ may not be fitted to generated images, we refine these estimates by minimizing the difference between the rendered masks from the model and the mask obtained using SAM.

We use a soft-rasterizer to render a silhouette $\hat{m} = \Pi(\hat{M}, \theta)$ where Π represents the renderer. The goal is to align this rendered mask $\hat{m}$ with the SAM-generated segmentation mask m'.

To refine the D-SMAL pose parameters θ, we minimize the following objective:

$$\theta^* = \min_{\theta} \|m' - \hat{m}\|. \tag{7}$$

The refinement process ensures that the D-SMAL model better captures the detailed geometry and appearance, adapting the parameters θ until the rendered and target masks are closely aligned.

Acknowledgements This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-II220612, Geometric and Physical Commonsense Reasoning based Behavior Intelligence for Embodied AI, and No.RS-2020-II201336, Artificial Intelligence Graduate School Program (UNIST)).

Funding Open Access funding enabled and organized by Ulsan National Institute of Science and Technology (UNIST)

Data Availability The GART dataset (Lei et al., 2024) is available at https://github.com/JiahuiLei/GART. The DFA dataset (Luo et al., 2022) is available at https://github.com/HaiminLuo/Artemis. The internet-sourced dog dataset (Dognomics2023, 2023) is available at https://www.dognomics.com/breed-guide.

References

AlBahar, B., Saito, S., Tseng, H.-Y., Kim, C., Kopf, J., & Huang, J.-B. (2023) Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia*,

Bagautdinov, T., Wu, C., Simon, T., Prada, F., Shiratori, T., Wei, S.-E., Xu, W., Sheikh, Y., & Saragih, J. (2021). Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG), 40*(4), 1–17.

Biggs, B., Boyne, O., Charles, J., Fitzgibbon, A., & Cipolla, R. (2020) Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop. In *ECCV*,

Couairon, G., Verbeek, J., Schwenk, H., & Cord, M. (2023) Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR*,

Dognomics2023. Internet-source dataset. https://www.dognomics.com/breed-guide/, n.d.

He, Z., & Wang, T. (2023) Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM,

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

Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., & Nie, L. (2023a) Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. *arXiv preprint* arXiv:2312.02134,

Hu, S., Hong, F., Pan, L., Mei, H., Yang, L., & Liu, Z. (2023b) Sherf: Generalizable human nerf from a single image. In *ICCV*,

Huang, Y., Yi, H., Liu,W., Wang, H., Wu, B., Wang, W., Lin, B., Zhang, D., & Cai, D. (2023) One-shot implicit animatable avatars with model-based priors. In *ICCV*,

Jain, A., Tancik, M., & Abbeel, P. (2021) Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *ICCV*,

Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., & Poole, B. (2022) Zero-shot text-guided object generation with dream fields. In *CVPR*,

Jiang, W., Yi, K. M., Samei, G., Tuzel, O., & Ranjan, A. (2022) Neuman: Neural human radiance field from a single video. In *ECCV*,

Jun, H., & Nichol. A. (2023) Shap-e: Generating conditional 3d implicit functions. *arXiv preprint* arXiv:2305.02463,

Kearney, S., Li, W., Parsons, M., Kim, K.I., & Cosker, D. (2020) Rgbd-dog: Predicting canine pose from rgbd sensors. In *CVPR*,

Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023) 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4), July URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, & W.-Y., et al. Segment anything. *arXiv preprint* arXiv:2304.02643, 2023.

Kocabas, M., Chang, J.-H. R., Gabriel, J., Tuzel, O., & Ranjan, A. (2023) Hugs: Human gaussian splats. *arXiv preprint* arXiv:2311.17910,

Kwon, Y., Kim, D., Ceylan, D., & Fuchs, H. (2021). Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems, 34*, 24741–24752.

Lei, J., Wang, Y., Pavlakos, G., Liu, L., & Daniilidis, K. (2024) Gart: Gaussian articulated template models. In *CVPR*,

Li, R., Tanke, J., Vo, M., Zollhöfer, M., Gall, J., Kanazawa, A., & Lassner, C. (2022) Tava: Template-free animatable volumetric actors. In *ECCV*,

Li, Z., Litvak, D., Li, R., Zhang, Y., Jakab, T., Rupprecht, C., Wu, S., Vedaldi, A., & Wu, J. (2024) Learning the 3d fauna of the web. In *CVPR*,

Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y., & Lin, T.-Y. (2023) Magic3d: High-resolution text-to-3d content creation. In *CVPR*,

Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., & Theobalt, C. (2021). Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG), 40*(6), 1–16.

Liu, M., Xu, C., Jin, H., Chen, L., Xu, Z., & Su, H., et al. (2023a) One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint* arXiv:2306.16928,

Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., & Vondrick, C. (2023b) Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*,

Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., & Wang, W. (2023c) Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint* arXiv:2309.03453,

Luo, H., Xu, T., Jiang, Y., Zhou, C., Qiu, Q., Zhang, Y., Yang, W., Xu, L., & Yu, J. (2022) Artemis: Articulated neural pets with appearance and motion synthesis. *ACM Trans. Graph.*, 41(4),

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020) Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*,

Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., & Chen, M. (2022) Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint* arXiv:2212.08751,

Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., & Bao, H. (2021a) Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*,

Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., & Zhou, X. (2021b) Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*,

Poole, B., Jain, A., Barron, J.T., & Mildenhall, B. (2023) Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*,

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021) Learning transferable visual models from natural language supervision. In *ICML*,

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022) Hierarchical text-conditional image generation with clip latents,

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022) High-resolution image synthesis with latent diffusion models. In *CVPR*,

Rüegg, N., Tripathi, S., Schindler, K., Black, M.J., & Zuffi, S. (2023a) Bite: Beyond priors for improved three-d dog pose estimation. In *CVPR*,

Rüegg, N., Zuffi, S., Schindler, K., & Black, M. J. (2023). Barc: Breed-augmented regression using classification for 3d dog reconstruction from images. *IJCV, 131*(8), 1964–1979.

Saito, S., Yang, J., Ma, Q., & Black, M.J. (2021) Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*,

Sinha, S., Shapovalov, R. Reizenstein, J., Rocco, I., Neverova, N., Vedaldi, A., & Novotny, D. (2023) Common pets in 3d: Dynamic new-view synthesis of real-life deformable categories. In *CVPR*, 2023.

Tang, J., Ren, J., Zhou, H., Liu, Z., & Zeng, G. (2023) Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint* arXiv:2309.16653,

Weng, C.-Y., Curless, B., Srinivasan, P. P., Barron, J. T., & Kemelmacher-Shlizerman, I. (2022) HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *CVPR*,

Wu, S., Li, R., Jakab, T., Rupprecht, C., & Vedaldi, A. (2023) MagicPony: Learning articulated 3d animals in the wild. In *CVPR*,

Xu, J., Wang, X., Cheng, W., Cao, Y.-P., Shan, Y., Qie, X., & Gao, S. (2023) Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *CVPR*,

Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Chang, H., Ramanan, D., Freeman, W.T., & Liu, C. (2021) Lasr: Learning articulated shape reconstruction from a monocular video. In *CVPR*,

Yang, G., Vo, M., Neverova, N., Ramanan, D., Vedaldi, A., & Joo, H. (2022) Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*,

Yang, G., Wang, C., Reddy, N.D., & Ramanan, D. (2023) Reconstructing animatable categories from videos. In *CVPR*,

Yang, Z., Zhou, M., Shan, M., Wen, B., Xuan, Z., Hill, M., Bai, J., Qi, G.-J., & Wang, Y. (2024) Omnimotiongpt: Animal motion generation with limited data. In *CVPR*,

Yao, C.-H., Hung, W.-C., Li, Y., Rubinstein, M., Yang, M.-H., & Jampani, V. (2023) Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In *CVPR*,

Ye, J., Wang, P., Li, K., Shi, Y., & Wang, H. (2023) Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. *3DV*,

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., & Wang, O. (2018) The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*,

Zuffi, S., Kanazawa, A., Jacobs, D., & Black, M.J. (2017) 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*,

Zuffi, S., Kanazawa, A., Berger-Wolf, T., & Black, M.J. (2018) Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *ICCV*,