

Every Dog Has Its Day, Probably: A Balanced Synthetic Benchmark and Probabilistic Modeling for 3D Dog Pose Estimation

Joo Young Choi^{1*}, Wonkwang Lee^{1,2*†}, Juhyeong Seon¹, and Gunhee Kim²

¹ KRAFTON AI, Seoul, Republic of Korea

{jychoi, wonkwang.lee, juhyeong.seon}@krafton.com

² Seoul National University, Seoul, Republic of Korea

{wonkwang.lee, gunhee}@snu.ac.kr



Fig. 1: We present a large-scale synthetic benchmark (4.2M images with exact 3D GT) for 3D dog pose estimation, generated by combinatorially sampling from libraries of pose (70K), shape (500), and texture (1,000) covering the broad range of viewpoints, articulations, and body proportions underrepresented in existing benchmarks.

Abstract. 3D dog pose estimation is hindered by two limitations in existing benchmarks: *distributional bias* toward canonical viewpoints, stable poses, and narrow morphologies, and *alignment degradation* between images and 3D annotations. We instead synthesize a large-scale balanced benchmark (4.2M images with exact 3D ground truth) by decoupling pose, shape, and texture into independent libraries (70K retargeted poses, 500 breed shapes, 1,000 appearance maps) and rendering directly to guarantee pixel-perfect alignment and broad view coverage. However, this balanced distribution exposes a challenge previously masked by biased data: non-canonical combinations of pose, shape, and viewpoint create multimodal 3D ambiguities that deterministic regression cannot resolve. We therefore propose a flow-based probabilistic framework that models the conditional distribution of 3D poses given an image, producing multiple plausible hypotheses under ambiguity while converging to precise estimates for clear views. Experiments on StanfordExtra, Animal3D, and our synthetic benchmark demonstrate that our method, trained solely on synthetic data, outperforms state-of-the-art approaches (79.9 vs. 33.2 PCK on the synthetic benchmark). Dataset available here.

Keywords: 3D Dog Pose Estimation · Synthetic Data · Flow Matching

* Equal contribution. †Corresponding author.

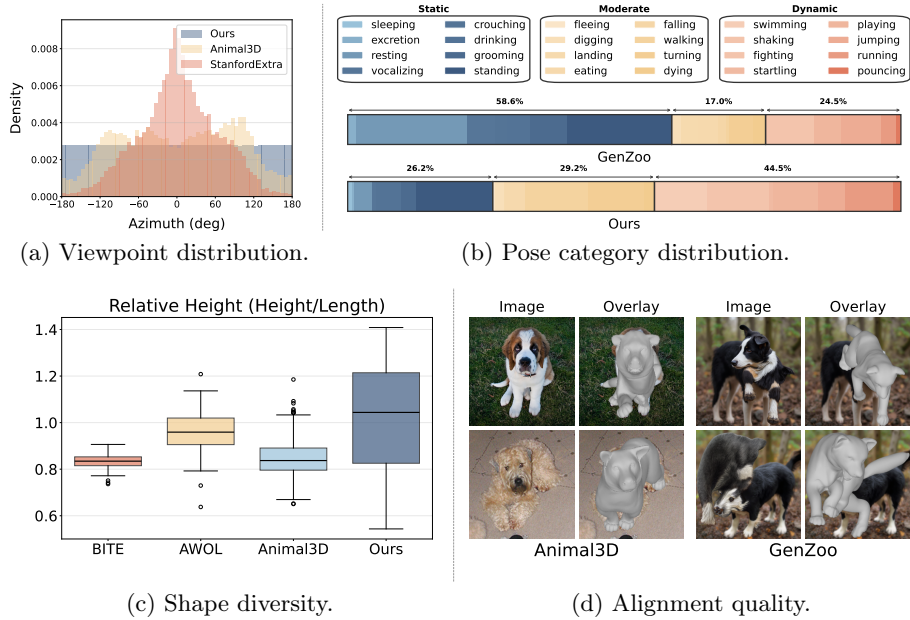


Fig. 2: Analysis of existing 3D dog pose benchmarks. (a) Existing viewpoint distributions [4, 34] cluster around a narrow range of canonical views; ours is uniform. (b) Existing pose distributions [20, 23] are largely limited to static motions; ours covers diverse articulations from static to dynamic actions. (c) Existing shape distributions [23, 34, 40] concentrate in a narrow band of the valid morphological spectrum [0.5, 1.5] [33]; ours covers [0.54, 1.41]. (d) Existing benchmarks rely on either AbS-fitted pseudo-GT [34] or diffusion-rendered images [20], both of which degrade image-to-3D alignment.

1 Introduction

Animal pose estimation forms the basis of a growing range of applications, from behavioral analysis [19, 22] and veterinary diagnostics [3, 39] to quadruped robotics [21, 38]. Among animals, dogs, our most common companion, pose a particularly demanding challenge: their vast morphological diversity, ranging from Corgis to Great Danes, coupled with the complex articulation of four limbs occupying overlapping depth planes, introduces self-occlusions and ambiguities that far exceed those found in human pose estimation [4, 23, 41].

Yet existing benchmarks for 3D dog pose estimation exhibit systematic limitations rooted in their shared reliance on early Analysis-by-Synthesis (AbS) approaches [4, 23, 24] (fitting a parametric 3D model [23, 40, 41] to 2D keypoint labels or detections), which is reliable only under canonical viewpoints, stable poses, and high keypoint visibility. As illustrated in Fig. 2, the resulting biases propagate through subsequent benchmarks [4, 20, 34, 40] as two intertwined limitations: **(i) Distributional bias:** (a) viewpoint coverage in the *images* is concentrated on canonical angles, while (b) pose diversity in the *3D annotations* is confined to stable or static motions and (c) shape diversity to a narrow morphological range; and **(ii) Alignment degradation:** (d) 3D annotations

become systematically less accurate for the images depicting non-canonical scenarios underrepresented by (i). These limitations reflect a fundamental issue of data *quality*, not *quantity*, and cannot be resolved by scaling data volume alone.

To break this cycle, we decouple data generation from real-image fitting entirely. Our key idea is to decompose dog geometry and appearance into three independent factors (pose, shape, and texture), each sourced without 2D-supervised curation, and combinatorially compose them to synthesize training data with pixel-perfect 3D alignment. Specifically, we build three independent libraries (70K poses from retargeted motion data covering static to dynamic actions, 500 breed-specific shapes and 1,000 textures leveraging text-to-image and image-to-textured-mesh models [14]), each curated for broader coverage than the canonical regime of prior benchmarks. Combinatorially sampling across these libraries with stratified viewpoints yields 4.2M images paired with exact 3D ground truth, with pixel-perfect alignment guaranteed by direct rendering.

This balanced distribution exposes a challenge previously masked by biased data: diverse non-canonical conditions (unusual viewpoints, complex articulations, and varied morphologies) create multimodal 3D pose ambiguities that deterministic regression struggles to resolve. We address this with a flow-based probabilistic framework that models the conditional distribution of 3D poses given an image, producing multiple plausible hypotheses under ambiguity while converging to precise estimates for clear views.

Trained solely on synthetic data, our method outperforms all baselines on StanfordExtra [4] and Animal3D [34] real-world benchmarks, with the gap widening substantially on our balanced synthetic test set (79.9 vs. 33.2 PCK). Controlled ablations confirm that each dataset component addresses a distinct failure mode (viewpoint bias, pose complexity, and shape diversity), while the probabilistic formulation yields further gains under ambiguity.

Our contributions are:

1. A large-scale synthetic benchmark (4.2M images with exact 3D ground truth) constructed from pose, shape, and texture libraries, each independently sourced to achieve broad coverage, whose combinatorial sampling provides balanced distributions that existing datasets lack.
2. A flow-based probabilistic estimator that models the multimodal 3D pose posterior, generating multiple plausible hypotheses under self-occlusion rather than a single deterministic prediction.
3. State-of-the-art results on StanfordExtra (88.3 PCK) and Animal3D (93.0 PCK) using only synthetic training data; ablations confirm that each dataset component addresses a distinct failure mode and that probabilistic regression yields further gains under ambiguity.

2 Related Work

3D Quadruped Pose and Shape Estimation. Parametric mesh models lie at the core of 3D quadruped reconstruction: SMAL [41] introduced a learned articulated model, DSMAL [23] added breed-level shape parameters, and SMAL+ [40]

broadened the shape space with additional scans. Building on these representations, Analysis-by-Synthesis (AbS) methods (WLDO [4], BARC [24], BITE [23]) fit the mesh to images via 2D keypoint reprojection, but the optimization is under-constrained under self-occlusion, producing poses that reproject correctly yet remain anatomically implausible. Recent methods (GenZoo [20], AniMer+ [1]) regress 3D pose from synthetic images, but rely on AbS-fitted parameters as supervision and diffusion-based rendering, inheriting distributional biases while introducing pixel-level misalignment at non-canonical viewpoints (Fig. 2-d). Our work breaks this cycle by sourcing pose, shape, and texture from independent libraries (each designed for broad coverage) without relying on 2D-supervised AbS fitting, and rendering with pixel-perfect alignment.

Synthetic Data for 3D Dog Pose Estimation. While marker-based motion capture [10, 30, 37] provides accurate 3D trajectories, it is confined to controlled settings with few subjects. Several rendering-based benchmarks address this gap: SyDog [27–29] renders Unity rigs with 2D keypoint annotations, DigiDogs [31] extracts 3D skeletons from GTA V, and RGBT-Dog [6] builds a custom parametric model from mocap of 11 dogs. However, each benchmark is individually limited in diversity: poses from at most 16 subjects, shapes and textures from at most 8 fixed meshes or a PCA-based linear subspace fitted to ≤ 11 scans, too few to span the full range of canine morphology and appearance. Moreover, none adopt a standard parametric model such as SMAL/DSMAL, yielding inconsistent joint definitions and incompatibility with the broader literature. We address both issues by building on DSMAL, ensuring unified topology and enabling combinatorial composition of pose, shape, and texture with broad per-factor coverage (70K motions, 500 shapes, 1,000 textures).

Probabilistic Pose Estimation. Deterministic regression predicts a single pose per input, an assumption that breaks down under heavy occlusion where multiple 3D configurations explain the same 2D observation. In human pose, multi-hypothesis approaches, such as mixture density networks [13], normalizing flows (ProHMR [11]), and diffusion-based methods (D3DP [26]), have shown improved accuracy under occlusion. These formulations have not been applied to animal pose, where ambiguity is more pronounced: quadruped articulation causes limb joints to be self-occluded from most viewpoints, and broad morphological variation adds further uncertainty. We adopt Rectified Flow [12, 15, 16] with an MMDiT [8] architecture and show that probabilistic modeling is equally critical for 3D dog pose and that its benefit is amplified by balanced training data, without which multiple hypotheses introduce noise rather than useful diversity.

3 Method

We present a unified framework for robust 3D dog pose estimation. Our approach addresses two fundamental challenges: the lack of large-scale 3D data that is both distributionally diverse and precisely aligned, and the inherent ambiguity of pose estimation under the occlusions that arise across diverse poses and viewpoints. Section 3.1 addresses the first challenge by constructing a synthetic benchmark, and Section 3.2 addresses the second via probabilistic pose estimation.

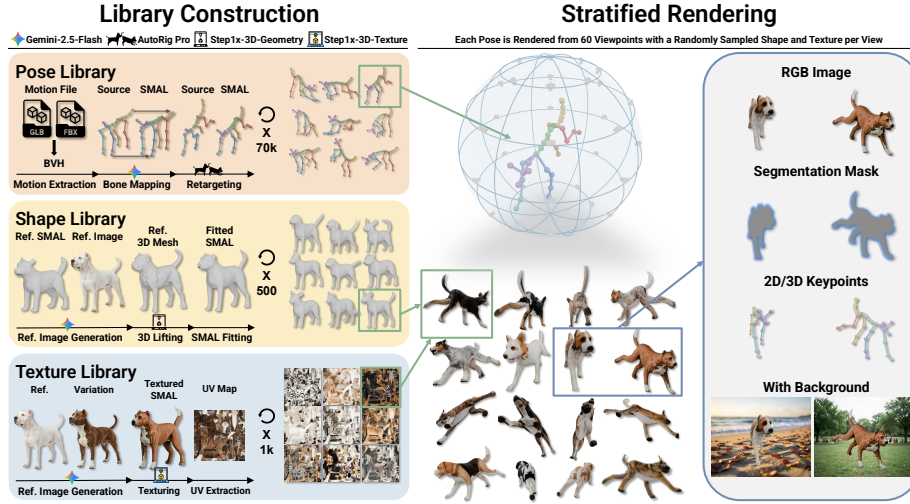


Fig. 3: Overview of our Synthetic Data Generation Pipeline. (Left) We construct three independent libraries (pose, shape, and texture) to ensure broad coverage and reduce distributional bias. From top to bottom: a pose library of 70K poses is built by retargeting artist-created motion data; a shape library of 500 breed-specific meshes and a texture library of 1,000 appearance maps are constructed leveraging text-to-image (Gemini-2.5-Flash) and image-to-textured-mesh (Step1x3D [14]) models. (Right) For each pose, 60 cameras are placed on a uniformly sampled viewing sphere, and a random shape–texture combination is drawn per view. Each rendered sample is paired with its segmentation mask, 2D/3D keypoint ground truth, and camera parameters; backgrounds are composited on the fly during training.

3.1 Constructing a Balanced Synthetic Benchmark

In this work, we decouple the representation of dog geometry and appearance into pose, shape, and texture, constructing an independent library for each factor so that their combinatorial sampling yields multiplicative coverage. Additionally, the benchmark must provide exact 3D pose ground truth for every sample, without relying on manual annotation or fitting-based estimation. To satisfy both requirements, we adopt a parametric mesh model as the basic building block of our benchmark, as we described below.

Parametric Mesh Model. We adopt DSMAL [23], which extends SMAL [41] to better capture morphological variance across dog breeds. This model is a differentiable function $\mathcal{M}(\theta, \beta_{\text{body}}, \beta_{\text{limbs}}) \rightarrow \mathcal{V}$, where $\mathcal{V} \in \mathbb{R}^{3889 \times 3}$ are the mesh vertices. The parameters comprise pose $\theta \in \mathbb{R}^{35 \times 6}$ (root orientation and 34 joint rotations in 6D rotation representation), body shape $\beta_{\text{body}} \in \mathbb{R}^{30}$ controlling global deformation, and limb proportions $\beta_{\text{limbs}} \in \mathbb{R}^9$ for legs, tail, ears, and torso. We denote $\{\beta_{\text{body}}, \beta_{\text{limbs}}\}$ as β hereafter.

DSMAL satisfies both requirements outlined above. Its parametric decomposition into pose θ and shape β enables independent control over each factor and thus combinatorial sampling. Its fixed topology further enables texture reuse: a

texture created for one mesh transfers directly to any other pose-shape combination, making texture an independent third factor. Finally, 3D keypoints can be read directly from vertex positions, providing exact 3D GT without manual annotation. The following subsections detail each library (pose, shape, and texture) with the full pipeline in Fig. 3.

Pose Library. As shown in Fig. 3 (left-top), we construct our pose library by retargeting 7K motion sequences from Sketchfab and our internal motion-capture and artist-created motion data onto the DSMAL skeleton. Since source files come in diverse formats (e.g., .bvh, .fbx, .glb/.gltf), we first convert all motions into a unified BVH representation. We then use Gemini-2.5-Flash to infer bone-to-bone semantic correspondences between each source skeleton and DSMAL joints from BVH metadata (e.g., `L_UpLeg` $\rightarrow$ `LeftHip`). Each motion is retargeted to DSMAL using Auto-Rig Pro [2], obtaining pose parameters θ per frame. All retargeted sequences are visually inspected, correcting erroneous bone mappings and filtering out implausible results.

Although this yields approximately 200K valid poses, directly using all would produce a skewed distribution: consecutive frames are temporally similar, and common poses (e.g. upright standing, neutral trot) recur across sequences, biasing the library toward overrepresented configurations. Thus, we further deduplicate in 6D rotation space, filtering out samples with cosine similarity above 0.99, yielding approximately 35K unique poses. Finally, we apply left-right mirroring to double the count to 70K unique poses.

The resulting library differs from prior benchmarks in two key aspects. First, it provides broader action coverage: as shown in Fig. 2(b), existing datasets are dominated by static poses (e.g. standing), whereas ours contains a higher proportion of dynamic actions (e.g. dying, playing, running). Second, because the source motions are authored directly in 3D and visually verified at the motion level, the resulting poses avoid the quality degradation that AbS-derived estimates suffer under non-canonical viewpoints, complex articulations, and self-occlusion. Full details of the pose library construction and its distributional analysis are provided in Secs. S3.1 and S4.2 of the supplementary material.

Shape Library. Accurately localizing joints in 3D often requires implicit knowledge of body proportions [23,24]. This is especially critical for self-occluded parts, where no direct 2D evidence is available and the model must rely on learned priors to infer 3D positions. Thus, we construct a shape library of 500 breeds from a purebred dog database [7] covering a broad range of the body-proportion space, as in Fig. 3 (left-middle).

First, we set the DSMAL pose θ to a *standard show pose* and render the mesh from a three-quarter viewpoint that maximizes visibility of all semantic parts (head, tail, and all four limbs). We then prompt Gemini-2.5-Flash-Image to convert this reference into a target breed while preserving the pose and viewpoint (e.g. `convert this dog into a Guatemalan Dogo while keeping the same pose and viewpoint`). These images are lifted to 3D meshes using Step1X-3D [14], a feed-forward image-to-3D model, with symmetric control enabled to ensure left-right symmetric meshes.

Since the resulting meshes possess arbitrary topologies incompatible with DSMAL, we recover shape parameters β by non-rigid registration of the DSMAL template to each target mesh, minimizing a combination of Chamfer distance, regularization, symmetry, and pose consistency losses. Full details are provided in Sec. S3.2 of the supplementary material.

We verify coverage using the *relative height* metric from the morphometric study of Sutter *et al.* [33], which measured 1,155 dogs across 109 AKC-recognized breeds. As shown in Fig. 2(c), our reconstructed shapes span a broad range of relative height values within the biologically plausible range reported in that study, whereas baseline shape sources concentrate in a narrow band.

Texture Library. As illustrated in Fig. 3 (left-bottom), we reuse the breed-specific images generated during shape library construction and additionally prompt Gemini-2.5-Flash-Image to synthesize diverse coat style variants (*e.g.* **change the coat pattern to brindle with everything kept unchanged**), yielding 1,000 reference images in total. Each image is then fed into the image-based texture synthesis pipeline of Step1X-3D [14] to produce a high-fidelity texture map on the canonical UV space of the DSMAL model. The resulting texture maps form the texture library, and since they share a common UV layout, each texture can be applied to any pose and shape configuration.

Stratified Rendering. As shown in Fig. 3 (right), we adopt a *stratified view-point sampling* strategy to avoid the canonical view bias found in prior datasets. We partition the viewing sphere into 60 regions ($4 \text{ azimuth} \times 5 \text{ elevation} \times 3 \text{ roll bins}$) and uniformly sample camera parameters within each bin. For each pose, we render all 60 viewpoints, independently sampling a random shape and texture from their respective libraries per view. Each sample is stored with its RGB image, segmentation mask, camera parameters, 3D keypoint coordinates, and their 2D projections under the corresponding camera, yielding 4.2 million samples with exact 3D ground truth. Since the rendered images contain no background, we separately pre-generate a pool of 1,000 diverse background images depicting plausible dog environments using Flux2 [5]. At training time, the dataloader on-the-fly composites each rendered foreground onto a randomly sampled background using the segmentation mask.

3.2 Probabilistic Pose Estimation via Rectified Flow

Our balanced dataset exposes a challenge masked by narrow benchmark coverage: ambiguity. Unlike curated datasets dominated by canonical views, balanced coverage over viewpoints, articulations, and morphologies naturally introduces diverse self-occlusions where a single 2D image may correspond to multiple plausible 3D poses. Deterministic regression models are ill-suited for these scenarios, as they cannot represent multimodal distributions. Fig. 4 illustrates our framework, detailed below.

Problem Formulation. To address this, given an input image $\mathbf{c}$, we formulate pose estimation as learning the conditional distribution $p(\mathbf{x}^{3D}|\mathbf{c})$ of the target $\mathbf{x}^{3D}$, which concatenates 3D coordinates of J keypoints $\in \mathbb{R}^{J \times 3}$ and weak-

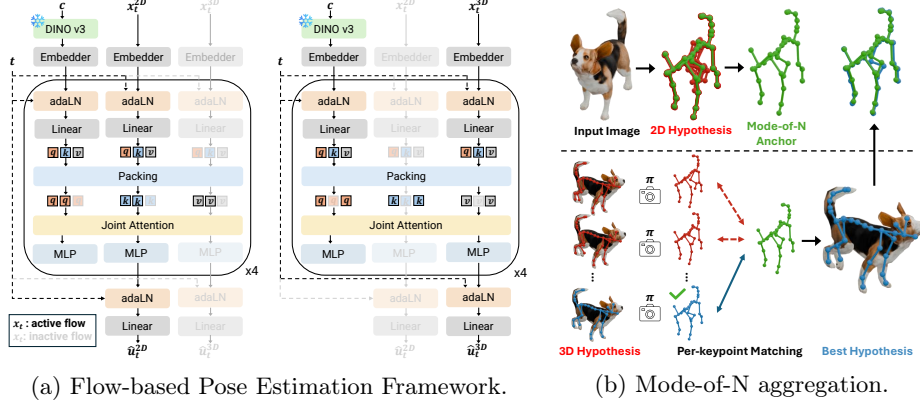


Fig. 4: Overview of our probabilistic pose estimation framework. (a) A frozen DINOv3 backbone encodes the input image into patch tokens that condition an MMDiT-based Rectified Flow network. The network operates in two modes: a 2D mode that activates image and 2D keypoint modules, and a 3D mode that activates image and 3D keypoint modules. (b) Mode-of-N aggregation: a per-keypoint mode is computed from the N 2D predictions to form a 2D reference pose, and the 3D hypothesis whose projection best matches it is selected as the final output.

perspective camera parameters (s, t_x, t_y) . In practice, we additionally learn a 2D target $\mathbf{x}^{2D} \in \mathbb{R}^{J \times 2}$ of projected keypoints within the same model: since our probabilistic formulation generates multiple 3D hypotheses, we use the 2D predictions as a reliable anchor for selecting the best 3D hypothesis at inference, following the principle introduced by D3DP [26] of leveraging 2D cues to aggregate 3D candidates.

Rectified Flow Matching. To learn this distribution, we adopt Rectified Flow [12, 16, 18], which constructs straight-line ODE trajectories from a noise prior $p_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the target distribution. For each task, let $(\mathbf{c}, \mathbf{x}_1)$ denote an image and its ground-truth target (i.e., $\mathbf{x}^{3D}$ or $\mathbf{x}^{2D}$), and $\mathbf{x}_0 \sim p_0$ a noise sample. We define the noisy state $\mathbf{x}_t$ as a linear interpolation between the two:

$$\mathbf{x}_t = t\mathbf{x}_1 + (1 - t)\mathbf{x}_0, \quad t \in [0, 1] \quad (1)$$

The velocity field of this path is simply $u_t(\mathbf{x}_t) = \mathbf{x}_1 - \mathbf{x}_0$. We train a neural network $v_\theta(\mathbf{x}_t, t, \mathbf{c})$ to predict this velocity field by minimizing the flow matching objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{t \sim \mathcal{U}(0,1), \mathbf{x}_0, \mathbf{x}_1, \mathbf{c}} [\|v_\theta(\mathbf{x}_t, t, \mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2] \quad (2)$$

This formulation is applied independently to both tasks, yielding two losses $\mathcal{L}^{2D}$ and $\mathcal{L}^{3D}$ that are summed during training.

Network Architecture. We parameterize v_θ using the MMDiT [8] architecture, popularized by Stable Diffusion 3, with three modality modules: an image module, a 2D keypoint module, and a 3D keypoint module, each with its own

weights. The input image $\mathbf{c}$ is encoded by a frozen DINOv3-base [32] backbone into patch tokens, while each component of the noisy state $\mathbf{x}_t$ is tokenized by its respective modality module. In each transformer block, each module independently projects its tokens into queries, keys, and values, which are then concatenated across active modalities for a single joint attention operation. In each forward pass, only the image module and the task-relevant modality modules are active: the image and 2D keypoint modules for the 2D task, and the image and 3D keypoint modules for the 3D task.

Inference. At inference, we solve the ODE $d\mathbf{x}_t/dt = v_\theta(\mathbf{x}_t, t, \mathbf{c})$ from $t = 0$ to $t = 1$ using an Euler solver. Since different noise initializations $\mathbf{x}_0 \sim p_0$ may converge to different modes of the posterior, a single sample does not guarantee the best solution. We therefore draw N independent samples to generate N 2D and N 3D hypotheses from their respective flows, leveraging the diversity of the probabilistic model and the higher accuracy of the 2D flow to select the best 3D output. Specifically, we compute a per-keypoint mode from the N 2D predictions to form a 2D reference pose, then select the 3D hypothesis whose projection best matches it. We refer to this aggregation protocol as **Mode-of-N** in Section 4.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on two complementary real-world benchmarks alongside our synthetic test set. **StanfordExtra** [24] provides 2D keypoint annotations across 120 dog breeds (12K images) but lacks 3D ground truth, restricting evaluation to 2D reprojection metrics. **Animal3D** [34] provides pseudo-3D annotations obtained via AbS fitting across multiple animal species; we restrict evaluation to the 20 canine breeds, but its curation retains only samples where we find AbS succeeds (50 images). We complement these with our synthetic test set (**Synthetic**), comprising 30K image–3D ground truth pairs across 6 held-out subjects (2 per size class: small, medium, large), each rendered from 60 uniformly sampled viewpoint bins per pose. These subjects are held-out rigged mesh animations sourced from the same online repositories as our pose library; poses are deduplicated following the same protocol described in Section 3.1. Unlike training data, the 3D models are rendered directly with their original textures without DSMAL fitting, and 3D keypoints are extracted from the original rig; we evaluate only the joints corresponding to DSMAL keypoints.

Baselines. We compare against three representative prior methods that employ deterministic regression of SMAL parameters (*i.e.* pose (θ) and shape (β)) from a single image. **BITE** [23] is an AbS method that optimizes the parameters at test time by minimizing the reprojection error of predicted 2D keypoints. **GenZoo** [20] is a direct regression model trained on synthetic image–3D pairs generated by rendering AbS-fitted meshes via diffusion models. **AniMer+** [1] is a hybrid: it trains on 3D ground truth when available and on 2D-annotated real images otherwise, by regressing SMAL and camera parameters and supervising their 2D projection against annotated 2D keypoints. Additionally, we evaluate a

deterministic variant of our model, **Ours (Det.)**, which shares the same MMDiT architecture but replaces the noisy input tokens with learnable query tokens, producing a direct prediction in a single forward pass without iterative denoising.

Metrics. We report Mean Per-Joint Position Error (**MPJPE**) [9] and Percentage of Correct Keypoints at threshold 0.10 (**PCK**) [36]. For 3D evaluation (Animal3D, Synthetic), we use Procrustes-aligned variants to factor out differences in camera assumptions across methods. For deterministic baselines, we extract 3D keypoints from their predicted SMAL meshes. For our probabilistic model, we report two aggregation protocols: **Mode-of-N** (Section 3.2) and **Best-of-N**, which selects the 3D hypothesis closest to the ground truth. Mode-of-N evaluates practical single-prediction quality without oracle access, while Best-of-N measures the model’s ability to cover the true pose among its hypotheses.

Implementation Details. We train our depth-4 MMDiT-based estimator for 500k iterations with a batch size of 256 on 4 NVIDIA A100 GPUs. The input images are resized to 256×256 . During inference, we sample $N = 20$ hypotheses independently for Mode-of-N and Best-of-N.

4.2 Comparison with State-of-the-Art

Table 1 compares our method against baselines across all three benchmarks. Our probabilistic model (Mode-of-N and Best-of-N) achieves the best results in every setting, consistently outperforming all baselines in both MPJPE and PCK.

Sim-to-Real Transfer. Despite being trained solely on synthetic data, our method generalizes well to real-world benchmarks. On StanfordExtra, Ours (Mode) outperforms BITE (85.57 vs. 82.88 PCK), which was trained on the StanfordExtra training set itself. Similarly, on Animal3D, Ours (Mode) surpasses AniMer+ (87.70 vs. 84.55 PCK), which used both the StanfordExtra and Animal3D training sets.

Synthetic Benchmark. The performance gap widens substantially on our synthetic test set, which covers a balanced distribution of viewpoints and a broad range of poses. The strongest baselines, AniMer+ and GenZoo, reach only 32–33 PCK, while even Ours (Det.) achieves 55.06, suggesting that training data distribution may play a significant role. Ours (Mode) further improves to 64.97, demonstrating the additional benefit of probabilistic modeling over deterministic regression under the same training data. A fine-grained stratified analysis by viewpoint and visibility is provided in Fig. S5 of the supplementary material, further confirming that baselines degrade on non-canonical viewpoints and under self-occlusion, while our method maintains consistent performance.

Qualitative Comparison. Fig. 5 presents qualitative results on six in-the-wild images downloaded from the web, spanning a range of viewpoints, visibility levels, and action categories from static to dynamic (standing, mounting, rolling over, leaping, belly down, and sleeping). For each case, we compare our prediction against AniMer+, the strongest baseline in the quantitative evaluation (Table 1); full comparisons with all baselines are provided in the supplementary material. For better visibility, we additionally show mesh reconstructions: while baselines can directly synthesize SMAL meshes by design, we recover only

Table 1: Comparison with state-of-the-art on two real-world benchmarks (StanfordExtra, Animal3D) and our synthetic benchmark. All variants of our method are trained exclusively on synthetic data. Our probabilistic model outperforms all baselines across all settings, with the gap widening on the balanced synthetic benchmark.

Method	StanfordExtra		Animal3D		Synthetic	
	MPJPE ↓	PCK ↑	MPJPE ↓	PCK ↑	MPJPE ↓	PCK ↑
BITE	0.1405	82.88	0.2738	63.03	0.2040	12.22
GenZoo	0.1588	77.50	0.1653	70.03	0.1210	33.21
AniMer+	0.1537	78.04	0.1289	84.55	0.1165	32.73
Ours (Det.)	0.1368	83.62	0.1878	69.56	0.0842	55.06
Ours (Mode)	0.1321	85.57	0.1120	87.70	0.0768	64.97
Ours (Best)	0.1211	88.30	0.0910	93.00	0.0606	79.88

SMAL shape parameters from our predicted 3D keypoints via test-time optimization, replacing the 2D detection supervision in BITE [23] with our 3D estimates. Even on a standing pose viewed from behind with all limbs visible, AniMer+ misplaces distal limb keypoints, suggesting limited understanding of body proportions. Under harder viewpoints (*e.g.* mounting, rolling over) or self-occlusion (*e.g.* leaping, belly down, sleeping), it defaults to a canonical standing pose, implying the limitation of canonical-biased training data and deterministic regression. Our method maintains accurate localization across all conditions, including challenging viewpoints and self-occlusion.

4.3 Ablation Study

The state-of-the-art comparison above involves methods that differ in training data, backbone architecture, and output representation, and the aggregate metrics can mask specific failure modes. To disentangle these factors, we conduct a controlled ablation study on our synthetic benchmark, progressively replacing each dataset component and comparing regression types under matched conditions. Each configuration is named **VPSR-X**: the first four slots encode the dataset (**V**iew, **P**ose, **S**hape, **R**endering), each set to either **O** (ours) or **B** (baseline), and **X** encodes the regression type, either **D** (deterministic) or **P** (probabilistic). All probabilistic models report Mode-of-N.

Since generating the full 4.2M dataset per configuration is prohibitively expensive, we use a reduced subset of 1K poses $\times$ 60 viewpoints (60K images) per configuration. For each component, the **O** variant uses our data while the **B** variant uses existing benchmarks. **V**: our uniform sphere sampling vs. Animal3D viewpoint distribution; **P**: 1K from our 70K library vs. 1K from Animal3D and GenZoo, both selected by clustering into 50 groups and sampling 20 per group; **S**: our 500 breeds vs. 500 shapes from Animal3D and GenZoo; **R**: our 1,000 textures rendered directly vs. ControlNet diffusion rendering following GenZoo. We organize the analysis into two groups: (1) progressive dataset degradation under deterministic regression (**OOOO**-D $\rightarrow$ **BBBB**-D), and (2) the effect of probabilistic regression on the endpoints (**OOOO**-P vs. **OOOO**-D and **BBBB**-P vs. **BBBB**-D).

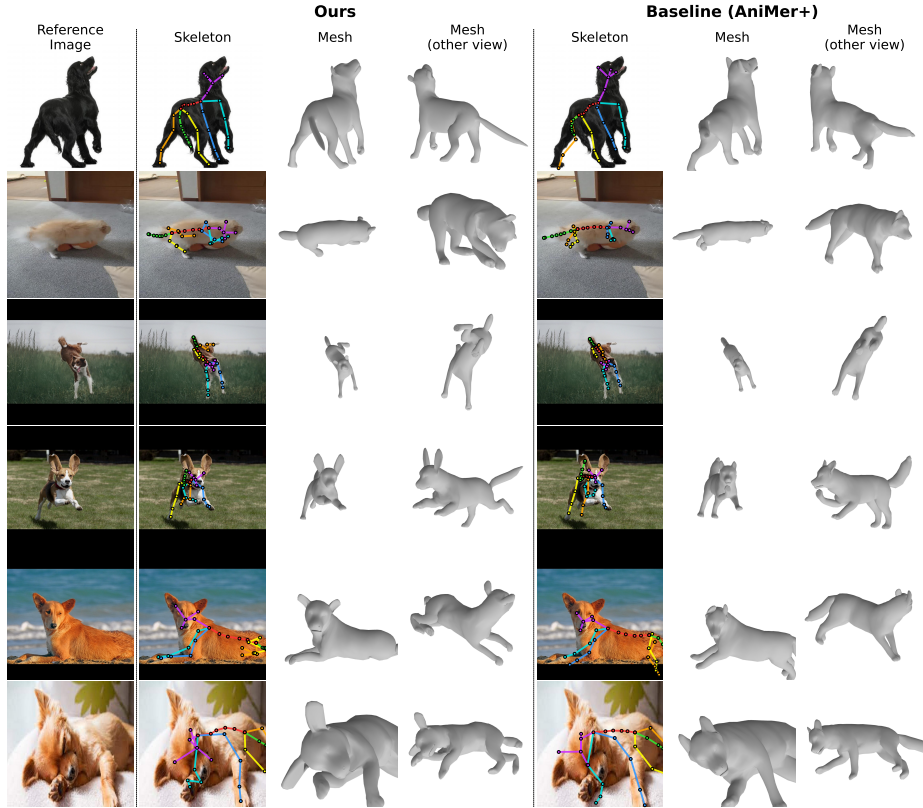


Fig. 5: Qualitative comparison with AniMer+ on six in-the-wild images. Each row shows, from left to right: input RGB, our results (3D keypoint overlay, mesh projection, rotated mesh view), and AniMer+ results in the same format. Cases span static to dynamic actions (standing, mounting, rolling over, leaping, belly down, sleeping) under diverse viewpoints and visibility conditions. Additional comparisons against all baselines (BITE, GenZoo, AniMer+) on the StanfordExtra and Animal3D benchmarks are provided in Figs. S1–S4 of the supplementary material.

Dataset Ablation ($\text{0000-D} \rightarrow \text{BBBB-D}$). Fixing regression to deterministic, we progressively replace each dataset component with its baseline counterpart; performance decreases monotonically with each replacement (Fig. 6). Each degradation is localized to the corresponding axis: replacing balanced viewpoints ($\text{0000-D} \rightarrow \text{B000-D}$) introduces viewpoint-dependent degradation in the azimuth plane; substituting retargeted motion data with baseline poses ($\text{B000-D} \rightarrow \text{BB00-D}$) causes a prominent drop in the most dynamic poses (Q4); replacing diverse shapes with mean-centric shapes ($\text{BB00-D} \rightarrow \text{BBB0-D}$) degrades performance on small-sized breeds; and switching to diffusion-based rendering ($\text{BBB0-D} \rightarrow \text{BBBB-D}$) produces the largest single drop across all axes, as it disrupts the pixel-to-geometry correspondence.

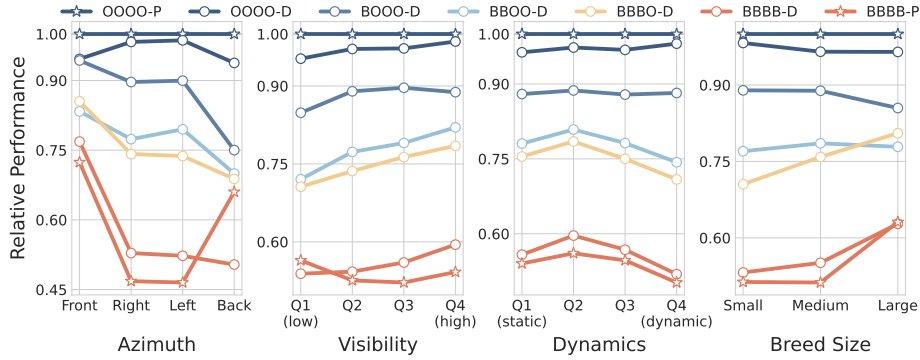


Fig. 6: Ablation Results. Relative PCK-PA@0.10 with respect to our full model (0000-P, Mode-of-N), stratified by azimuth, visibility, dynamics, and breed size.

Probabilistic vs. Deterministic. On our full dataset, 0000-P consistently outperforms 0000-D, with gains concentrated in ambiguous scenarios: low visibility (Q1), front/back azimuths, and large breeds. On the baseline dataset, however, BBBB-P underperforms BBBB-D, indicating that without sufficient distributional coverage, multiple samples introduce noise rather than meaningful diversity. This confirms that probabilistic regression and balanced training data are complementary: the formulation provides the mechanism to handle ambiguity, while distributional coverage supplies the knowledge to do so effectively.

Role of Visual Backbone. In the dataset ablation (0000-D $\rightarrow$ BBBB-D) and the probabilistic comparison (0000-P vs. BBBB-P), the model architecture and frozen DINOv3 backbone remain identical, yet performance varies substantially with training data alone. A similar trend appears in the cross-method comparison (Table 1), where baselines with different backbones exhibit the same viewpoint- and visibility-dependent degradation (see Fig. S5). These results consistently point to distributional coverage of training data as a key factor in performance.

4.4 Multi-Hypothesis Visualization

Fig. 7 demonstrates that our probabilistic model captures a meaningful multi-modal distribution over 3D poses under self-occlusion. The input shows a leaping dog from a near-frontal viewpoint, where both hind legs are entirely hidden behind the torso and the front legs nearly overlap in 2D.

To visualize this, we adopt a two-stage resampling strategy inspired by RePaint [17]. First, we draw $N = 64$ hypotheses via Mode-of-N and classify joints as *confident* (variance $< 0.1 \times$ bounding-box scale, mirroring PCK@0.10) or *unconfident*. Second, we resample by applying RePaint to our rectified flow: confident joints follow a deterministic flow from their Mode-of-N estimates, while unconfident joints explore diverse modes. In this example, the confident joints are the visible frontal parts (head, torso, front paws); the mesh overlay confirms an accurate Mode-of-N fit on these regions. The four rotated views from Re-

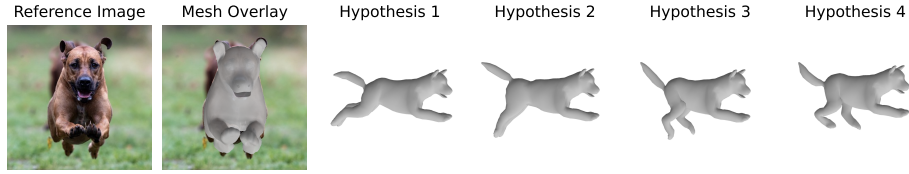


Fig. 7: Multi-hypothesis visualization via RePaint. From left to right: input image, Mode-of-N mesh overlay, and four hypotheses generated by fixing confident joints and resampling unconfident joints via RePaint. The overlay confirms an accurate Mode-of-N fit on visible parts (head, torso, front paws), while the rotated views reveal diverse yet plausible hind-leg configurations from the second-stage resampling.

Paint resampling show that the occluded hind legs take on distinct yet plausible configurations while confident joints remain fixed. This confirms that our model commits to precise estimates where visual evidence is available and maintains diverse, plausible alternatives where it is not.

5 Conclusion

We presented a balanced synthetic benchmark of 4.2 million images paired with exact 3D ground truth and a Rectified Flow-based probabilistic estimator, showing that the two are complementary: balanced data supplies distributional coverage, while probabilistic modeling provides the mechanism to handle the ambiguity that this coverage exposes. The resulting model, trained solely on synthetic data, achieves state-of-the-art results on two real-world benchmarks; we hope this progress contributes to downstream applications such as behavioral analysis, veterinary diagnostics, and quadruped robotics.

Limitations. Our training pipeline composites rendered foregrounds onto backgrounds without simulating depth interactions, so it handles self-occlusion but not environmental occlusion (*e.g.* a dog partially hidden behind furniture). Depth-aware compositing, as employed by SAM 3D [25] to synthesize natural occlusion from background objects, could address this gap. Additionally, our method is trained exclusively on synthetic data; a multi-phase human-and-model-in-the-loop training scheme [25, 35], where model predictions on real images are curated and fed back as training data in successive stages, could further close the sim-to-real gap. Finally, as a single-image method, we do not exploit temporal information from video, which could reduce ambiguity through motion cues. On the asset side, the DSMAL [23] model we build on inherits SMAL’s linear blend skinning and, unlike SMPL for humans, lacks pose-dependent corrective blend shapes, so extreme articulations can induce local mesh degeneracy (*e.g.* knee interpenetration under strong flexion or lumbar volume loss in upright sitting); extending DSMAL with such corrective shapes would mitigate this.

Acknowledgements. This research was conducted as part of the Sovereign AI Foundation Model Project(Data Track), organized by the Ministry of Science and ICT(MSIT) and supported by the National Information Society Agency(NIA), S.Korea. (Grant No. 2026-AIData-WII01).

References

1. An, L., Lyu, J., Lin, L., Cheng, P., Liu, Y., Tang, X.: AniMer+: Unified pose and shape estimation across mammalia and aves via family-aware transformer. IEEE TPAMI (2026)
2. Artell: Auto-rig pro. <https://www.lucky3d.fr/auto-rig-pro/doc/index.html> (2023), accessed: 2025-10-01
3. Barney, S., Dlay, S., Crowe, A., Kyriazakis, I., Leach, M.: Deep learning pose estimation for multi-cattle lameness detection. Scientific Reports (2023)
4. Biggs, B., Boyne, O., Charles, J., Fitzgibbon, A., Cipolla, R.: Who left the dogs out: 3D animal reconstruction with expectation maximization in the loop. In: ECCV (2020)
5. Black Forest Labs: FLUX.2: Frontier Visual Intelligence. <https://bfl.ai/blog/flux-2> (2025), accessed: 2026-01-01
6. Deane, J., Kearney, S., Kim, K.I., Cosker, D.: RGBT-Dog: A parametric model and pose prior for canine body analysis data creation. In: WACV (2024)
7. DogPack: Purebred dog breeds A-Z list. <https://www.dogpackapp.com/blog/dog-breeds/purebred/> (2025), accessed: 2025-10-01
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
9. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE TPAMI (2014)
10. Kearney, S., Li, W., Parsons, M., Kim, K.I., Cosker, D.: RGBD-Dog: Predicting canine pose from rgb-d sensors. In: CVPR (2020)
11. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: ICCV (2021)
12. Lee, W., Jeong, J., Moon, T., Kim, H.J., Kim, J., Kim, G., Lee, B.U.: How to Move Your Dragon: Text-to-Motion Synthesis for Large-Vocabulary Objects. In: ICML (2025)
13. Li, C., Lee, G.H.: Generating multiple hypotheses for 3d human pose estimation with mixture density network. In: CVPR (2019)
14. Li, W., Zhang, X., Sun, Z., Qi, D., Li, H., Cheng, W., Cai, W., Wu, S., Liu, J., Wang, Z., et al.: Step1X-3D: Towards high-fidelity and controllable generation of textured 3d assets. arXiv:2505.07747 (2025)
15. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
16. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: ICLR (2023)
17. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Re-Paint: Inpainting using denoising diffusion probabilistic models. In: CVPR (2022)
18. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: ECCV (2024)
19. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M.: DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience (2018)

20. Niewiadomski, T., Yiannakidis, A., Cuevas-Velasquez, H., Sanyal, S., Black, M.J., Zuffi, S., Kulits, P.: Generative Zoo. In: ICCV (2025)
21. Peng, X.B., Coumans, E., Zhang, T., Lee, T.W.E., Tan, J., Levine, S.: Learning agile robotic locomotion skills by imitating animals. In: Robotics: Science and Systems (2020)
22. 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., McKenzie-Smith, G.C., Mitelut, C.C., Diez Castro, M., D’Uva, J., Kislin, M., Sanes, D.H., Kocher, S.D., Wang, S.S.H., Falkner, A.L., Shaevitz, J.W., Murthy, M.: SLEAP: A deep learning system for multi-animal pose tracking. *Nature Methods* (2022)
23. Rüegg, N., Tripathi, S., Schindler, K., Black, M.J., Zuffi, S.: BITE: Beyond priors for improved three-D dog pose estimation. In: CVPR (2023)
24. Rüegg, N., Zuffi, S., Schindler, K., Black, M.J.: BARC: Learning to regress 3D dog shape from images by exploiting breed information. In: CVPR (2022)
25. SAM 3D Team, Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., Lin, A., Liu, J., Ma, Z., Sagar, A., Song, B., Wang, X., Yang, J., Zhang, B., Dollár, P., Gkioxari, G., Feiszli, M., Malik, J.: SAM 3D: 3Dfy anything in images. *arXiv:2511.16624* (2025)
26. Shan, W., Liu, Z., Zhang, X., Wang, Z., Han, K., Wang, S., Ma, S., Gao, W.: Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In: ICCV (2023)
27. Shooter, M., Malleson, C., Hilton, A.: SyDog: A synthetic dog dataset for improved 2d pose estimation. *arXiv:2108.00249* (2021)
28. Shooter, M., Malleson, C., Hilton, A.: MVSyDog: A multi-view 3d dog pose dataset for advancing 3d pose estimation. In: ACM SIGGRAPH Eur. Conf. Visual Media Production (2023)
29. Shooter, M., Malleson, C., Hilton, A.: SyDog-Video: A synthetic dog video dataset for temporal pose estimation. *IJCV* (2023)
30. Shooter, M., Malleson, C., Hilton, A.: Benchmarking monocular 3d dog pose estimation using in-the-wild motion capture data. In: CVPRW (2024)
31. Shooter, M., Malleson, C., Hilton, A.: DigiDogs: Single-view 3d pose estimation of dogs using synthetic training data. In: WACV Workshops (2024)
32. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. *arXiv:2508.10104* (2025)
33. Sutter, N.B., Mosher, D.S., Gray, M.M., Ostrander, E.A.: Morphometrics within dog breeds are highly reproducible and dispute Rensch’s rule. *Mammalian Genome* (2008)
34. Xu, J., Zhang, Y., Peng, J., Ma, W., Jesslen, A., Ji, P., Hu, Q., Zhang, J., Liu, Q., Wang, J., et al.: Animal3D: A comprehensive dataset of 3d animal pose and shape. In: ICCV (2023)
35. Yang, X., Kukreja, D., Pinkus, D., Sagar, A., Fan, T., Park, J., Shin, S., Cao, J., Liu, J., Ugrinovic, N., Feiszli, M., Malik, J., Dollar, P., Kitani, K.: SAM 3D Body: Robust full-body human mesh recovery. *arXiv:2602.15989* (2026)
36. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. *IEEE TPAMI* (2013)
37. Zhang, H., Starke, S., Komura, T., Saito, J.: Mode-adaptive neural networks for quadruped motion control. *ACM TOG* (2018)

38. Zhao, L., Luo, Z., Han, Y., Zhang, J., Chen, Y., Liu, Y., Lu, P.: Learning aggressive animal locomotion skills for quadrupedal robots solely from monocular videos. *npj Robotics* (2025)
39. Zhu, H., Salgırlı, Y., Can, P., Atılgan, D., Salah, A.A.: Video-based estimation of pain indicators in dogs. In: *ACII* (2023)
40. Zuffi, S., Black, M.J.: AWOL: Analysis without synthesis using language. In: *ECCV* (2024)
41. Zuffi, S., Kanazawa, A., Jacobs, D., Black, M.J.: 3D menagerie: Modeling the 3D shape and pose of animals. In: *CVPR* (2017)