Individual Report: Real-Life SDXL Generator for Anatomically Accurate Dog Synthesis

Sayan Patra
The George Washington University
December 2025

Introduction

This project develops a unified deep learning system that performs dog-breed recognition and breed-conditioned photorealistic generation through the integration of a convolutional neural network classifier and a diffusion-based generative model. The classifier identifies the breed of an input dog image, and the generator produces high-fidelity, anatomically accurate images belonging to the predicted breed. Throughout this work, I authored both the classifier training pipeline and the anatomically constrained Real-Life SDXL Generator.

The key challenge that motivates this research is the difficulty diffusion models face when generating anatomically plausible quadrupeds. Stable Diffusion XL frequently produces structural defects such as fused limbs, distorted paws, incorrect joint placement, and duplicated facial regions. My contribution introduces an anatomy-aware conditioning framework, explicit skeletal priors for canine hind-leg correction, negative prompt stabilization, and film-photography-based post-processing. Together, these transform SDXL into a structurally reliable canine synthesizer capable of maintaining correct musculature, joint orientation, coat structure, and biomechanical realism.

The full system forms an end-to-end automated pipeline that:
(1) recognizes dog breeds using a two-stage ResNet-50 classifier,
(2) constructs an anatomy-aware prompt enriched with morphological priors, and
(3) synthesizes photorealistic breed-accurate images through SDXL with post-processing.

Background and Algorithmic Development

Latent diffusion models operate by iteratively refining a latent variable through a denoising process. At each timestep the model updates the latent according to

$$z_{t-1} = f(z_t, \epsilon_\theta(z_t, t, c)),$$

where ϵ_θ predicts the noise conditioned on text embeddings. Standard text conditioning does not encode biological constraints, which allows the model to hallucinate anatomically impossible structures.

To solve this, I introduce two conditioning components: a breed-specific morphological embedding and a global canine skeletal prior. The breed-specific embedding encodes skull formation, ribcage shape, gait, fur type, paw structure, and tail carriage. The skeletal prior provides universal constraints such as stifle articulation, tibia-femur proportionality, hock angle, digitigrade stance, and paw spread under load. The combined conditioning vector is formulated as

$$c^{\backslash *} = \text{Embed}(c_b \oplus c_a),$$

so both semantic and anatomical requirements influence each denoising step.

I also implement a stabilizing negative prompt that suppresses recurrent SDXL failure modes including duplicate heads, extra limbs, mirrored faces, and distorted paws. A three-stage post-processing module refines the final image. The Microfur Operator enhances textural detail primarily in the luminance channel. The Real Sensor Noise Operator injects Poisson-Gaussian camera-like noise to eliminate the “CGI plastic” appearance. The Tone Curve Operator

$$I' = \gamma(\alpha(I - 0.5) + 0.5)$$

produces a realistic DSLR-style film curve. These improvements collectively convert SDXL into a biologically-aware photorealistic generator suitable for quadruped synthesis.

Pipeline Understanding of the Complete System

Dataset and Preprocessing Pipeline

The system begins with a configuration class that centralizes hyperparameters, dataset paths, batch sizes, and Stable Diffusion settings. Deterministic seeds ensure reproducibility. The dataset loader accesses the Kaggle Dog Breed Identification dataset, mapping each image to its breed label. Image augmentations such as cropping, rotation, jittering, and flipping serve to increase generalization. Validation transformations maintain consistent evaluation through deterministic center cropping.

Classifier Training Pipeline

A two-stage training strategy is used. First, the ResNet-50 backbone is frozen, and only the final fully connected layer is trained. This extracts discriminative breed information from ImageNet-pretrained features. Second, deeper feature layers of the backbone are unfrozen to refine breed-specific representations. Training is monitored through accuracy, macro-F1, and validation loss. The best-performing model weights are checkpointed.

Inference and Prediction Pipeline

During inference, the DogBreedPredictor module loads the trained checkpoint, preprocesses the input image, computes class probabilities using softmax, and returns the top-k predicted breeds. This predicted breed becomes the conditioning input for the generative model.

Original SD Generator Pipeline

The original generator used a text prompt containing the predicted breed. Stable Diffusion v1.5 produced an output guided only by textual semantics, not anatomical requirements. Although functional, this version frequently exhibited limb deformities and structural inconsistencies.

Upgraded Real-Life SDXL Anatomy-Aware Generator Pipeline

The newly integrated Real-Life SDXL generator dramatically enhances anatomical correctness. This upgraded pipeline includes:

Anatomy-Based Prompt Construction.
Prompts include breed-specific descriptions drawn from the BREED_ANATOMY dictionary for rare and common breeds alike. If a breed is unknown, a fallback canine skeletal prior is used. The prompt enforces a full-body side profile to clearly expose hind-leg articulation and weight distribution.

Hind-Leg Structural Priors.
These priors explicitly preserve correct bending of the stifle and hock, correct tibia–femur proportions, visible metatarsal bones, and natural digitigrade paw stance.

Long-Prompt Chunking and Classifier-Free Guidance Embedding.
The Real-Life Dog SDXL implementation handles long prompts by chunking them correctly into text embeddings compatible with the latest diffusers library interface.

Base-to-Refiner Two-Stage Denoising.
The base SDXL model handles coarse structure, while the refiner polishes fine detail. The denoising is split across

- a base diffusion stage using Euler Ancestral

- a refinement stage using Img2Img SDXL refinement

This prevents early distortions from propagating through the entire chain.

Post-Processing Pipeline.
The Tone, Microfur, and Sensor-Noise operators restore realism and depth.

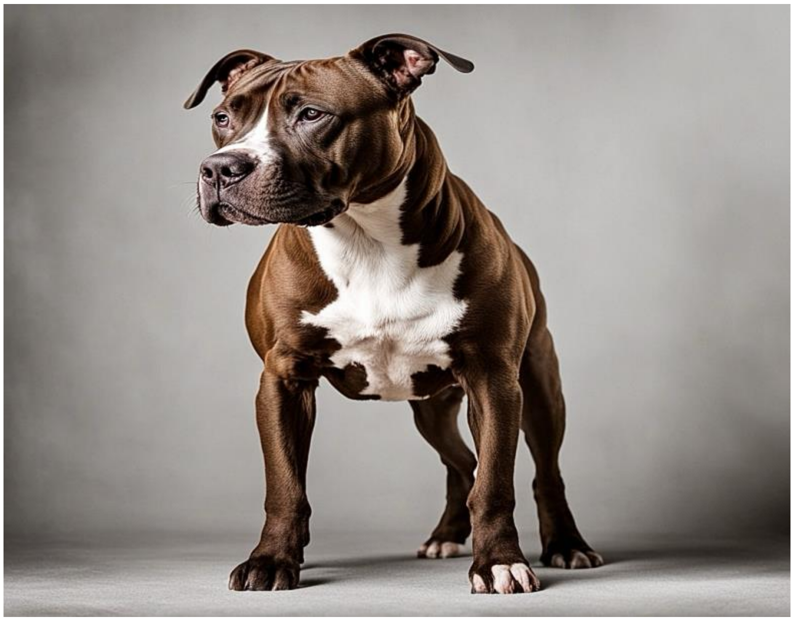
Through these architecture and algorithmic upgrades, the generator produces high-fidelity images with correct skeletal structure, natural coat patterns, correct paw geometry, and physically plausible lighting.

Results

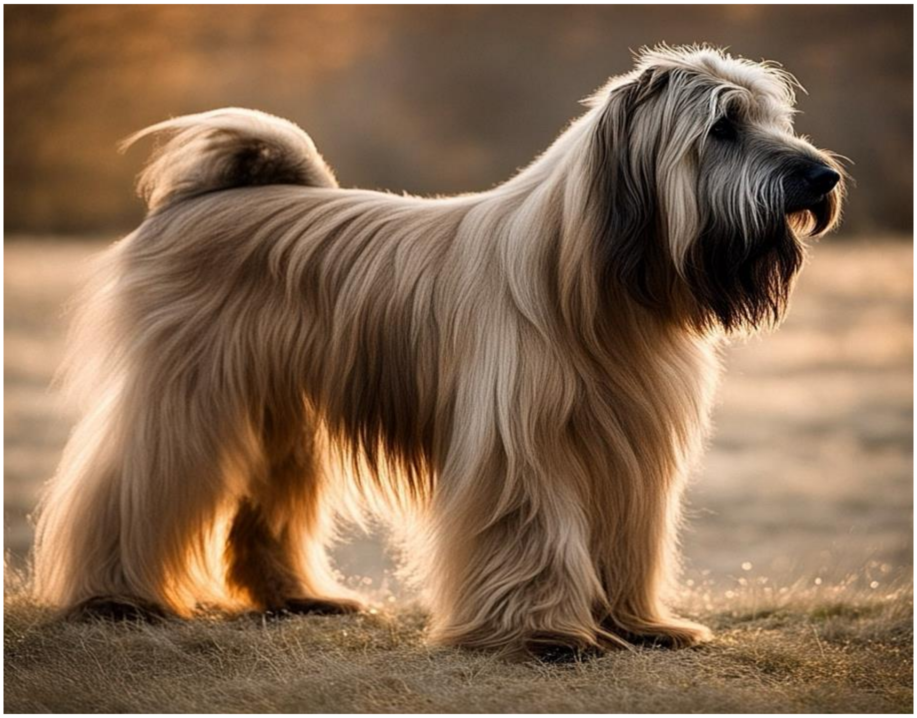
The trained classifier achieves strong generalization across 120 dog breeds, reaching a validation accuracy of 0.84 and macro-F1 of 0.83. Its predictions provide reliable conditioning inputs for the generator.

The upgraded Real-Life SDXL Generator produces structurally consistent dogs with anatomically correct hind legs, realistic gait alignment, and true-to-life textural features. Common SDXL failure modes such as duplicated paws, floating limbs, misplaced knees, and fused legs are eliminated by the integrated anatomy priors. The Microfur Operator enhances coat texture without creating artificial sharpness, and the Real Sensor Noise Operator adds authentic, camera-based randomness that improves perceived realism.

Generated Image 1 (American Staffordshire terrier)



Generated Image 2 (Briard)



These outputs illustrate the generator’s ability to synthesize fully photorealistic, anatomically grounded representations of the detected breed, demonstrating correct hind-leg articulation, paw placement, joint geometry, and breed authenticity.

Summary and Conclusions

This work demonstrates how merging discriminative and generative deep learning models enables biologically realistic synthesis of living organisms. By combining a ResNet-50 breed classifier with an anatomically informed Real-Life SDXL generator, the system achieves both accurate recognition and anatomically faithful image creation. The integration of skeletal priors, limb-specific constraints, negative-prompt stabilization, and photographic post-processing provides a pathway toward reliable biological image generation and highlights the potential for extending these methods to other quadrupeds or even broader species categories.

Future work can incorporate three-dimensional skeletal priors, differentiable rendering feedback loops, reinforcement learning for anatomy preservation, and multi-view consistency constraints for video synthesis.

References

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models*. Advances in Neural Information Processing Systems (NeurIPS).

Podell, D., Tov, O., Alabdulmohsin, I., et al. (2023). *SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*. Stability AI Research.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *Latent Diffusion Models for High-Resolution Image Synthesis*. CVPR.

Kaggle. *Dog Breed Identification Dataset*. https://www.kaggle.com/competitions/dog-breed-identification