Dog Breed Recognition & Image Generation System

Deep Learning - Group 6

Adam Stuhltrager
Junhua Deng
Sameer Batra
Sayan Patra

1. Introduction

This project develops an end-to-end dog-breed understanding pipeline that integrates modern deep learning for both recognition and generation. It tackles two complementary tasks: fine-grained dog-breed classification using transfer-learning-based ResNet models, and text-to-image dog generation using diffusion-driven Stable Diffusion. Because dog breeds often differ only in subtle visual details, combining a discriminative classifier with a generative model highlights a unified "understand and create" AI system capable of both identifying breeds and synthesizing realistic breed-specific imagery.

2. Related Works

Resnet-50 for Image Classification

ResNet-50 is part of the Residual Network family proposed by He et al. (2016), which introduced residual connections to mitigate the optimization challenges associated with training very deep convolutional networks. Its architecture consists of 50 layers built from bottleneck residual blocks, where each block compresses and expands feature dimensions through a 1×1 → 3×3 → 1×1 structure. This design significantly reduces computational cost while preserving representational depth, making it more efficient than earlier large CNNs such as VGG-16 and VGG-19.

Compared with traditional architectures, ResNet models offer several key advantages:

1. Improved gradient flow through identity shortcuts, enabling deeper and more stable training.
2. Higher accuracy on fine-grained classification tasks due to deeper hierarchical feature extraction.
3. Better parameter efficiency, as bottleneck blocks allow greater depth without proportional increases in model size.
4. Strong transfer learning performance, with ResNet-50 repeatedly outperforming models like AlexNet, VGG, and Inception in downstream applications.

ResNet-50 is chosen over shallower variants (e.g., ResNet-18 or ResNet-34) because fine-grained tasks such as dog-breed recognition require sensitivity to subtle differences in facial shape, coat texture, and posture—features more effectively captured by deeper architectures. At the same time, it offers a more favorable compute-to-performance ratio than deeper alternatives like ResNet-101 or ResNet-152, making it a practical choice for both training and real-time inference in an interactive application.

SDXL and LoRA

Stable Diffusion XL (SDXL) represents a significant advancement in latent diffusion models, providing higher-resolution synthesis, improved photorealism, and greater prompt controllability compared with earlier Stable Diffusion architectures. SDXL employs a two-stage architecture

consisting of a base model and a refinement model, enabling it to generate detailed images at 1024×1024 resolution. Its expanded text encoder capacity and multi-scale denoising pathways allow SDXL to capture complex semantic attributes, making it well suited for fine-grained subject control such as dog-breed–specific image generation.

To efficiently adapt large diffusion models like SDXL to new styles or visual domains, Low-Rank Adaptation (LoRA) has become a widely adopted technique. LoRA introduces a pair of low-rank matrices into pretrained model weights, enabling targeted fine-tuning with a minimal number of trainable parameters. This approach significantly reduces the computational cost of customization and avoids overfitting common in full-model finetuning. In practice, LoRA allows users to train or apply lightweight adapters representing specific artistic styles, camera types, or subject identities without modifying the underlying base model.

For this project, SDXL provides a strong foundation for high-quality generative outputs, while LoRA enables style specialization—together forming a controllable and computationally efficient generator for breed-specific and prompt-driven dog image synthesis.

3. Data Description

The dataset contains 10,222 dog images split into a training set and a test set, with each image identified by a unique filename. A csv file is provided for labeling the breed of dog in each image.

Link: https://www.kaggle.com/competitions/dog-breed-identification/overview

4. Exploratory Data Analysis

To gain an initial understanding of the data, we conducted an exploratory analysis on breed distribution. The dataset contains 10,222 dog images spanning 120 breeds, with an average of 85.18 samples per breed and a standard deviation of 13.30, indicating a moderately imbalanced distribution. The most represented breed is *scottish_deerhound* with 126 images, while *eskimo_dog* is the least represented with 66. The top five breeds by frequency are *scottish_deerhound*, *maltese_dog*, *afghan_hound*, *entlebucher*, and *bernese_mountain_dog*, each appearing over 110 times in the training set.

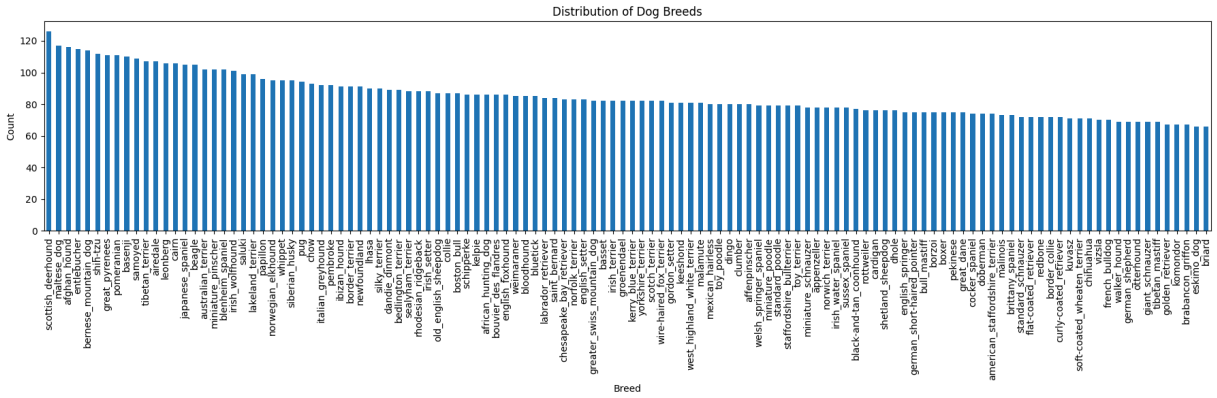

Fig.1: Distribution of Dog Breeds

5. Methodology

5.1 Pipeline Overview

The workflow consists of three integrated components designed to support dog-breed understanding and style-controlled image generation.

1) **Supervised Training of a Dog-Breed Classifier**
   The pipeline begins by training a ResNet-50 model on the original dog-breed dataset. After completing supervised training, the resulting .pth checkpoint is uploaded to the Hugging Face Hub to ensure reproducibility, model sharing, and seamless downstream integration.
2) **Style-Controlled Image Generation Using SDXL and LoRA**
   Stable Diffusion XL (SDXL) is then used to evaluate different LoRA adapters and prompt configurations for generating dog images in specific artistic or photographic styles. The chosen LoRA models and prompt templates are consolidated into a unified generator module to support consistent and controllable image synthesis.
3) **Deployment of an Integrated Streamlit Application**
   A Streamlit application was developed to integrate both the classifier and the generation modules within a unified interface. Users may upload an image to obtain a breed prediction from the ResNet-50 classifier or manually specify a breed if desired. For image synthesis, the application supports multiple LoRA-based styles—manga, anime, pastel anime, and pixel art—which are defined directly in the configuration and executed through the stylized SDXL generator. In contrast, the realistic style is handled by an independent SDXL-based realistic generator that operates outside the config-driven LoRA framework. Together, these components enable both classifier-guided and user-controlled breed-conditioned image generation across diverse visual styles.

5.2 Classification Module (ResNet-50)

Our training script fine-tunes a ResNet-50 model on the dog-breed dataset using a standard supervised learning setup. The dataset is loaded from *labels.csv* and the training images, with breed labels mapped to numerical indices. Images are preprocessed through resizing, random resized cropping, horizontal flipping, and normalization for training, while validation uses deterministic resizing and center cropping for consistent evaluation.

A pretrained ResNet-50 (ImageNet-1K) serves as the backbone, with the final classification layer replaced to match the number of breeds. Training is conducted for 15 epochs, using a batch size of 32, cross-entropy loss, and the AdamW optimizer with a learning rate of 1e-4 and weight decay of 1e-4. A Cosine Annealing learning-rate scheduler is applied across the training duration. The script automatically creates an 80/20 train–validation split and reports epoch-level training and validation metrics using tqdm progress bars.

This configuration provides a balanced setup that supports stable convergence while limiting computational cost, making it suitable for prototyping and downstream application integration.

5.3 Generative Module (SDXL + LoRA)

Our generator is built on the Stable Diffusion XL (SDXL) latent diffusion framework and integrates Low-Rank Adaptation (LoRA) to impose a specific artistic style. The module loads the SDXL base model and inserts LoRA weights into the model's cross-attention layers, where low-rank updates modulate attention transformations and introduce stylistic characteristics with minimal computational overhead.

Prompt and LoRA configuration followed a style-specific design strategy to ensure both semantic consistency and controllable stylistic variation. Each generation style corresponds to a predefined LoRA module and a tailored prompt template, allowing consistent conditioning across breeds. The selection of LoRA models was based on their demonstrated ability to impose distinct aesthetic transformations—such as manga line art, anime illustration, pastel rendering, or pixel-art stylization—while maintaining compatibility with the SDXL diffusion backbone.

For each style, the text prompt was constructed to include (1) an explicit breed placeholder to anchor semantic content, (2) descriptors enforcing correct canine anatomy, (3) stylistic tokens aligned with the corresponding LoRA domain (e.g., "inked lineart" for manga, "pastel colors" for soft-anime styles). Negative prompts were also style-dependent and were designed to suppress recurrent failure modes such as duplicated tails, anatomical distortion, or unwanted colorization in monochrome styles.

LoRA scaling values were chosen empirically to balance stylistic strength against structural fidelity; for example, manga and pixel-art LoRAs adopt moderate scaling (0.8–0.9) to avoid excessive abstraction, while anime LoRAs employ stronger scaling (≈1.5) to achieve the desired aesthetic intensity. Together, these components provide a reproducible prompt–LoRA pairing scheme that yields consistent, style-coherent dog images across all supported breeds.

Unlike the LoRA-driven styles defined in the configuration, the realistic style is provided through an independent SDXL generator module rather than the config system. It provides a separate two-stage base-and-refiner workflow, enhanced negative prompts, and camera-based post-processing filters for improved anatomical accuracy and photographic fidelity. In this way, the realistic generator extends the configuration framework by introducing a non-LoRA pathway that produces high-detail, naturalistic dog images.

5.4 Streamlit App Integration

A two-stage interactive application was implemented using Streamlit to integrate dog-breed classification with controllable image generation. The interface is organized into two functional tabs: Classify and Generate—each corresponding to a separate model pipeline. All core models and assets are loaded through cached resource functions to ensure reproducibility and minimize repeated initialization cost.

The classification tab enables users to upload an image, which is processed by a pretrained ResNet-50 dog-breed classifier. The classifier returns the top-k predicted breeds with associated confidence scores, and the highest-confidence prediction can be automatically transferred to the generation tab for downstream synthesis. Results are displayed with a structured UI element that highlights the predicted label and its confidence.

The generation tab provides a unified interface for synthesizing dog images in multiple visual styles. Users may select a breed from the predefined label set or manually input a custom breed. Style selection determines whether the system uses a LoRA-based SDXL generator or a realistic base-plus-refiner SDXL pipeline, with corresponding hyperparameters exposed through an advanced settings panel. The application dynamically routes requests to either the stylized generator or the realistic generator and performs inference according to user-specified guidance scale, spatial resolution, number of denoising steps, and optional random seed. Generated images are displayed immediately and can be exported via a built-in download function.

6. Experiments and Results

6.1 Resnet-50 Classifier

Evaluation

	Macro	Micro	Weighted
Precision	0.8508	0.8508	0.8601
Recall	0.8461	0.8508	0.8508
F1-score	0.8435	0.8508	0.8508
Accuracy			0.8508

The ResNet-50 classifier demonstrates strong overall performance across all evaluation metrics. Macro-averaged precision, recall, and F1-score are approximately 0.85, indicating that the model performs consistently across classes even when treating each breed equally, regardless of frequency. Micro metrics are identical to the overall accuracy (0.8508), showing that the model maintains stable performance when weighting predictions by sample count. The weighted precision is slightly higher (0.8601), suggesting that the model performs particularly well on more common breeds. Overall, the results show that ResNet-50 achieves balanced and reliable classification performance across the 120-dog-breed dataset.

Example use:
Fig. 2 predicted breed: Border_collie
Confidence: 0.99321448802948

Fig. 3 predicted breed: vizsla
Confidence: 0.9691827297210693

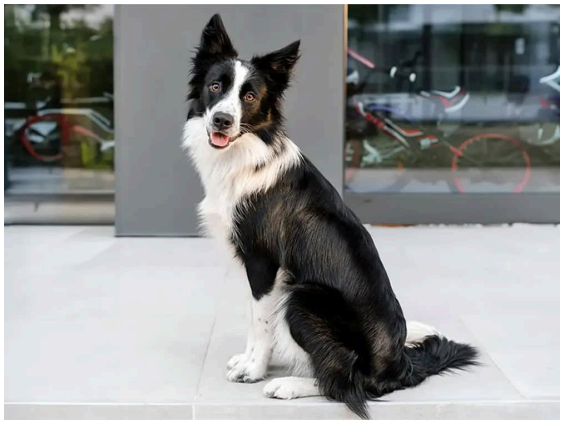

Fig. 2: Border Collie

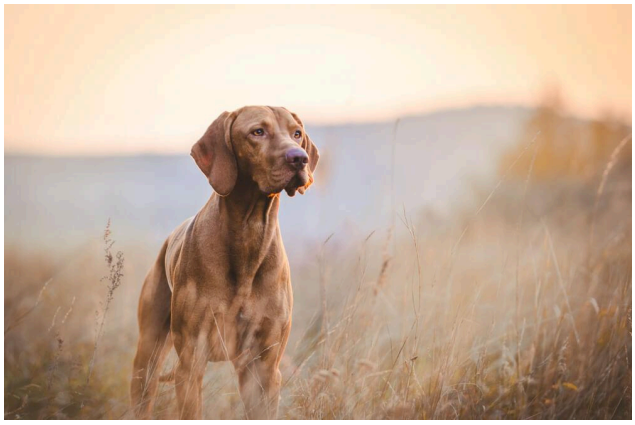

Fig. 3: Vizsla

6.2 Image Generator

The SDXL + LoRA framework successfully generated high-quality dog images across five stylistic domains: Realistic, Manga, Anime, Pastel Anime, and Pixel Art. The Realistic LoRA preserved anatomical fidelity and photorealistic texture, producing outputs with natural lighting and correct breed morphology. The Manga LoRA yielded clean monochrome line art with stable contours and stylized motion features, while the Anime LoRA produced vibrant color illustrations with expressive facial features and smooth shading. The Pastel Anime LoRA further softened the color palette, generating images with gentle lighting and a painterly, atmospheric aesthetic. In contrast, the Pixel Art LoRA produced low-resolution, grid-aligned outputs consistent with retro 16-bit game sprites. Across all five styles, the generator maintained strong breed recognizability with minimal structural artifacts, demonstrating the model's ability to adapt to diverse stylistic constraints without compromising semantic accuracy.

Example use:

Fig. 4:
Black and white side view of a **Boxer** dog, accurate canine anatomy, single visible tail, one tail only, proper proportions, full body in frame, natural limb spacing, dynamic pose, consistent perspective, shonen jump manga style, screentone shading, inked lineart, high contrast, speed lines, impact frame, dramatic action.

Fig. 5:
Beautiful illustration of a **Chihuahua** dog, anime style, vibrant colors, detailed fur, expressive eyes, soft lighting, peaceful scene, high quality.

Fig. 6:
Cute **Yorkshire Terrier** dog, pastel colors, soft anime style, gentle lighting, dreamy atmosphere, detailed fur, masterpiece, best quality.

Fig. 7:
Pixel art of a **Pug** dog, 16-bit style, retro gaming aesthetic, clean pixels, vibrant colors, side view sprite, detailed shading

Fig. 8:
RAW photo of a **Shetland Sheepdog** dog, full body side view, DSLR, 85mm lens, natural lighting, correct canine anatomy, four legs visible, proper joint angles, photorealistic, high detail, no stylization

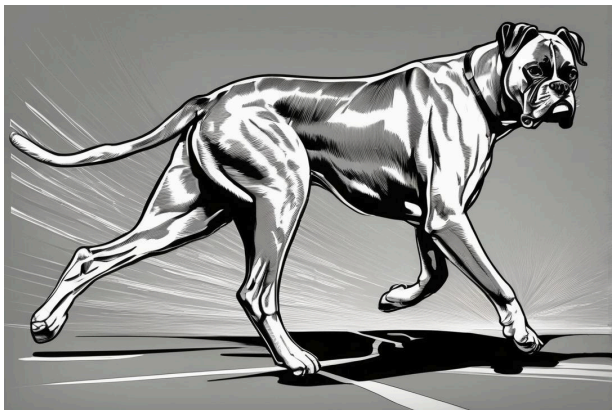

Fig. 4: Manga

Fig. 5: Anime

Fig. 6: Pastel Anime

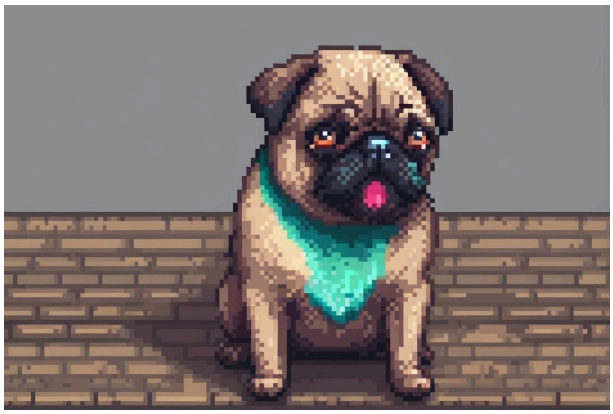

Fig. 7: Pixel Art

Fig. 8: Realistic

7. Discussion

This project demonstrates the feasibility of integrating a ResNet-50 dog-breed classifier with an SDXL-based generative model to create an end-to-end pipeline for breed recognition and multi-style image synthesis. The classifier achieved stable performance across the 120 breeds included in the training dataset, with macro-level metrics suggesting relatively balanced behavior despite class imbalance. Micro precision, recall, and F1-score matched the overall accuracy, reflecting consistent aggregate performance in a single-label multiclass setting. However, an important limitation is that the classifier can only recognize breeds present in the original label set; many real-world dog breeds fall outside these categories and therefore cannot be identified by the model, reducing its applicability beyond the closed-set dataset.

The SDXL + LoRA generator produced visually coherent outputs across the five supported styles—realistic, manga, anime, pastel anime, and pixel art—showing that SDXL's latent space preserves sufficient semantic structure to maintain breed identity while accommodating diverse stylistic transformations. Nevertheless, the generation pipeline exhibits structural vulnerabilities typical of diffusion models. SDXL occasionally produces incorrect limb or tail counts, distorted anatomy, or unstable body proportions, especially under strong stylistic conditioning or dynamic poses. These issues highlight the absence of explicit anatomical constraints in diffusion architectures and the difficulty of achieving consistent morphological accuracy even when negative prompts and LoRA scaling are applied.

The integrated Streamlit app demonstrates the practical value of pairing discriminative and generative models, allowing users to classify a dog image and immediately generate stylized outputs of the corresponding breed. While effective as a proof of concept, high-resolution SDXL inference remains computationally expensive, limiting real-time deployment. Overall, this project illustrates both the strengths and limitations of combining supervised classification with diffusion-based generation, pointing toward future improvements in dataset coverage, classifier generalization, structural priors for diffusion models, and optimization for interactive applications.

8. Conclusion

This project integrates a ResNet-50 dog-breed classifier with an SDXL + LoRA generator to create a unified system for breed recognition and multi-style image synthesis. The classifier performs reliably on the 120 labeled breeds, while the generative pipeline produces coherent outputs across five distinct visual styles. Although the system is limited by the classifier's closed-set breed coverage and occasional structural inconsistencies in SDXL-generated images, the results demonstrate the effectiveness of combining discriminative and generative models within an interactive application. This work provides a practical foundation for future improvements in classification generalization, anatomical control in diffusion models, and deployment efficiency.

9. References

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 770–778).

Podell, D., Tov, O., Alabdulmohsin, I., et al. (2023). *SDXL: Improving latent diffusion models for high-resolution image synthesis*. Stability AI Technical Report.

Hu, E. J., Shen, Y., Wallis, P., et al. (2021). *LoRA: Low-rank adaptation of large language models*. arXiv preprint arXiv:2106.09685.