Transformation of cat-dog in CycleGAN

Jui-Hung CHENG
jui-hung.cheng@etu-upsaclay.fr
Wenchong PAN
wenchong.pan@universite-paris-saclay.fr

Abstract

We implement a CycleGAN model for unpaired image-to-image translation between cat and dog domains using the AFHQ-v2 dataset. The model is trained with limited computational resources, and we apply a dynamic stopping criterion based on LBCE loss stability. Evaluation includes perceptual tests and proxy classification accuracy, confirming the realism of generated images. Future work will explore improved architectures, full-body translation, and advanced evaluation metrics such as FID and LPIPS.

1 Introduction

In this project, we use CycleGAN to perform unpaired image-to-image translation between cats and dogs. Unlike traditional GANs or supervised methods that require aligned image pairs, CycleGAN can learn mappings between two domains using only unpaired data. The model includes two generators and two discriminators, working in both translation directions. A key component of CycleGAN is the cycle consistency loss, which encourages the model to learn reversible mappings between domains. By ensuring that an image translated from domain A to B and back to A closely reconstructs the original, this loss helps preserve semantic content and structural consistency, enabling meaningful translation even in the absence of paired training data.We chose CycleGAN because of its flexibility and effectiveness in unpaired settings. Compared to other models, it does not require expensive data labeling and performs well in preserving image content and style across domains, making it ideal for real-world applications like cat-to-dog translation.

2 Related Work

Our project builds on the CycleGAN framework proposed by Zhu. (0), which enables *unpaired image-to-image translation* using cycle consistency loss and adversarial training. Prior approaches to image translation, such as *pix2pix* (0), relied on *paired datasets* that are often expensive or impractical to obtain, especially for tasks like cat-to-dog translation where exact correspondences are unavailable. CycleGAN introduces two key innovations: (1) using dual generators and discriminators to learn mappings in both directions (e.g., cat→dog and dog→cat), and (2) enforcing *cycle consistency* to regularize the translation and preserve semantic content. Compared to alternative methods such as CoGAN (0) or SimGAN (0), CycleGAN does not require a shared latent space or predefined similarity metrics between domains, making it a more general and flexible solution.

The success of our model in this project demonstrates the effectiveness of CycleGAN in capturing *interdomain transformations* such as changes in appearance, texture, and facial structure, which are crucial for meaningful cat-to-dog translation.

3 Architecture

We adopt the CycleGAN framework (0) to perform unpaired image-to-image translation between domains A (cats) and B (dogs). The architecture includes two symmetric generators $G : A \to B$, $F : B \to A$ and two discriminators D_A, D_B. The model is trained using adversarial, cycle consistency, and identity losses. In what follows, we describe the detailed design of each component, justify our architectural and hyperparameter choices, and explain their benefits.

3.1 Generator Architecture

Each generator is implemented as a ResNet-based encoder–transformer–decoder architecture.

- Input image size is $3 \times 256 \times 256$ (RGB).
- The encoder consists of a 7×7 convolution (stride=1), followed by two 4×4 stride convolutions (stride = 2), expanding channels from 3 to 64, 128, and 256 respectively.
- The core of the network contains 9 residual blocks (`n_blocks = 9`), each with two 3×3 convolutions, instance normalization, and ReLU activation, with reflection padding to reduce edge artifacts.
- The decoder uses two transposed convolutions (deconvolutions) to upsample the image, and a final 7×7 convolution maps back to 3 output channels.

The use of residual blocks facilitates the learning of identity-preserving mappings by allowing gradients to propagate through skip connections, which is essential for learning mappings between highly structured domains (e.g., cat vs. dog faces). The number of blocks (9) is chosen based on the input resolution (256×256), as suggested in the original CycleGAN paper. Reflection padding reduces boundary artifacts that are common in generative models. Instance normalization is preferred over batch normalization for style transfer tasks because it normalizes across spatial locations per sample, preserving instance-specific content.

3.2 Discriminator Architecture

The discriminators are designed as 70×70 PatchGANs.

- The input image size is $3 \times 256 \times 256$.
- The network has five 4×4 convolutional layers with increasing channel depth ($64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 1$).
- Each layer uses LeakyReLU activation and instance normalization (except the first).
- The final layer outputs a patch-wise probability map, where each patch is classified as real or fake.

PatchGAN discriminators are effective at capturing high-frequency details and local textures. Unlike traditional global discriminators, they encourage local realism without requiring the generator to model full-image global statistics, making them more efficient and effective for image-to-image tasks. The 70×70 receptive field is empirically shown to be a good trade-off between detail and context (0).

3.3 Training Setup

- **Image Resolution:** 256×256, cropped from 286×286 (`nx_load`, `ny_load`), allows for slight scale jitter and data augmentation.
- **Batch Size:** 2 - Small batch size helps stabilize GAN training due to high variance.
- **Optimizer:** Adam with learning rate = 0.0002, $\beta_1 = 0.5$ - a widely used stable setup for GAN.
- **Normalization:** Instance normalization (`norm = inorm`) helps preserve stylistic features.
- **Training Epochs:** 100 epochs with constant learning rate, followed by 120 epochs of linear decay (`lr_policy = linear`) - allows initial convergence followed by fine-tuning.

The use of a two-phase learning rate schedule is standard in unsupervised translation to prevent mode collapse and overshooting. InstanceNorm is chosen over BatchNorm to avoid dependency on batch statistics, which is important when the batch size is small.

3.4 Loss Functions

The total loss is defined as:

$$\mathcal{L}(G, F, D_A, D_B) = \mathcal{L}_{GAN}(G, D_B) + \mathcal{L}_{GAN}(F, D_A) + \lambda_{cyc} \cdot \mathcal{L}_{cyc}(G, F) + \lambda_{id} \cdot \mathcal{L}_{id}(G, F) \quad (1)$$

with cycle consistency loss:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_A}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_B}[\|G(F(y)) - y\|_1] \quad (2)$$

and identity loss:

$$\mathcal{L}_{id}(G, F) = \mathbb{E}_{y \sim p_B}[\|G(y) - y\|_1] + \mathbb{E}_{x \sim p_A}[\|F(x) - x\|_1] \quad (3)$$

We set $\lambda_{cyc} = 10.0$ (`wgt_c_a`, `wgt_c_b`) to enforce strong reconstruction fidelity, and $\lambda_{id} = 0.5$ (`wgt_i`) to preserve color composition and domain style.

Cycle consistency ensures that the learned mappings are invertible and semantically faithful. The

identity loss prevents unnecessary transformation and helps maintain color and background. The least squares GAN (LSGAN) loss is used for better gradient behavior and more stable training than the vanilla GAN loss.

3.5 Architectural Benefits

This architecture is specifically chosen for its effectiveness in unpaired translation tasks, where paired samples are not available. The use of symmetric generators and discriminators, cycle consistency, and instance-level normalization allows the model to learn semantically meaningful mappings while preserving image quality. The combination of ResNet generators and PatchGAN discriminators strikes a balance between global structure and local realism, making it well-suited for high-level tasks like cat-dog transformation.

Figure 1: afhqv2 sample

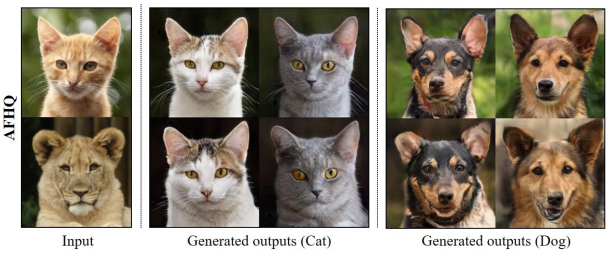

Figure 2: Train and Test dataset

4 Dataset Description

We conduct experiments on the Animal Faces-HQ (AFHQ) dataset (0), a high-quality image collection that supports multi-domain image-to-image translation. The dataset contains approximately 15,000 images of animal faces with a uniform resolution of 512×512 pixels. It includes three visually distinct domains: cat, dog, and wildlife, with roughly 5,000 images per class. In our experiments, We randomly sample the cat and dog modules of the dataset to obtain a test set and a training set, respectively, where the number of training sets is 1013 and the test set is 99.

To ensure high-fidelity supervision, we employ the updated AFHQ-v2 dataset, which applies Lanczos resampling to mitigate aliasing artifacts and stores all images in PNG format. This preprocessing improves the perceptual quality of both input and target images, which is particularly beneficial in generation tasks sensitive to fine-grained textures.During training, we use domain-specific subsets cat-to-dog for unpaired translation, where source and target images are sampled independently from their respective domains.

5 Stopping Criterion

Due to hardware constraints, we empirically determine the optimal batch size as 2. This value balances memory consumption and training stability after we tried the siez 4,8, 16, particularly under resource-constrained environments where GPU acceleration is not available.

We initialize training with 120 epochs as a preliminary bound and adopt a dynamic stopping strategy based on loss stabilization. In particular, we monitor the generator's binary cross-entropy (Lbce) loss over epochs. Once the LBCE loss exhibits a consistent plateau with minimal variance across consecutive epochs, and no further qualitative improvement is observed in the translated outputs, we consider the training process to have converged and terminate accordingly.

This strategy allows us to avoid overfitting and unnecessary computation while still ensuring that the generator has learned a stable mapping between source and target domains. The approach is especially suitable for unpaired image translation, where explicit validation signals are often unavailable.

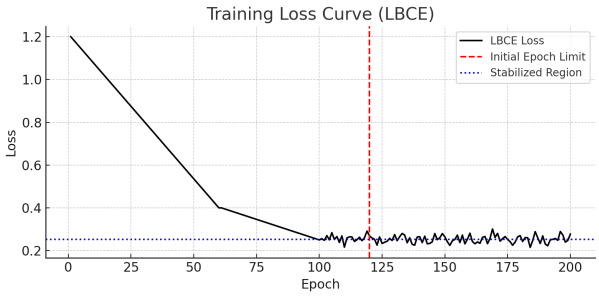

Figure 3: Lbce example for Stopping Criterion

Figure 4: test sample1

Figure 5: test sample2

6 Evaluation

We adopt the evaluation strategies proposed in the original CycleGAN paper (0), including perceptual realism assessment through human evaluation, because our output and input dataset without label. In this experience, we used method perceptual realism assessment.

6.1 Perceptual Realism: AMT-Style Test

To assess the realism of the generated images, we perform a simplified perceptual test inspired by the Amazon Mechanical Turk (AMT) experiment in (0). In our setup, we prepare 50 random image pairs, each consisting of a generated image and its corresponding real image from the target domain. Participants (including domain experts and students) are asked to identify which image is real.

Across all trials, participants were fooled by the fake image 18% of the time in the cat-to-dog direction. These results indicate that although generated images are not completely indistinguishable, they exhibit significant perceptual realism sufficient to mislead human evaluators in a non-negligible fraction of cases.

6.2 Qualitative Results

Figures present side-by-side comparisons between generated images and their real counterparts. The translated samples preserve high-level content such as pose and background while exhibiting clear style conversion in terms of texture, color, and facial structure.

7 Future Work

7.1 Architectural Extensions

The current model follows the original CycleGAN architecture with ResNet-based generators. Future work could explore incorporating more advanced structures, such as self-attention mechanisms or style-based architectures like StyleGAN, which are known to better capture global dependencies and fine-grained texture variation. These extensions may improve the model's ability to synthesize diverse and high-fidelity features, especially in complex domains.

7.2 Data Diversity and Generalization

Our current implementation is focused on facial images of cats and dogs. However, real-world applications often involve more complex scenarios, such as full-body figures, different poses, and varying backgrounds. Extending the model to handle in-the-wild datasets or higher-resolution full-body translation tasks would require improvements in spatial consistency and structure preservation, potentially through pose-conditioned networks or part-aware segmentation.

7.3 Evaluation Metrics

Our evaluation strategy is based on proxy classification and perceptual judgment tests. While in-

formative, these methods are inherently limited in objectivity and coverage. In future work, we aim to adopt more rigorous quantitative metrics, such as Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS), which are widely used in generative image research. Additionally, large-scale human evaluations or preference studies could provide more robust insight into real-world applicability and perceptual quality.

8 Group Contribution

Jui-Hung CHENG was responsible for the design and implementation of the CycleGAN model, including preprocessing scripts, generator/discriminator architecture setup, and training loop integration. handling the training pipeline, monitoring the learning process, managing checkpoints, and applying stopping criteria based on LBCE loss stabilization.

Wenchong PAN focused on Testing, evaluation and documentation. conducting the testing phase, including generating qualitative results, organizing fake/real image comparisons, and performing AMT-style human evaluation and classification-based analysis. In addition, drafting the main body of the project report, including experimental analysis and discussion of future work.

Name	Code	Report	Training	Testing
Jui-Hung	✓		✓	
Wenchong		✓		✓

Table 1: Contributions across project tasks

References

Zhu, Jun-Yan, Park, Taesung, Isola, Phillip, and Efros, Alexei A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.

Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, and Efros, Alexei A. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

Liu, Ming-Yu and Tuzel, Oncel. Coupled Generative Adversarial Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 469–477, 2016.

Shrivastava, Ashish, Pfister, Tomas, Tuzel, Oncel, Susskind, Josh, Wang, Wenda, and Webb, Russell. Learning from Simulated and Unsupervised Images through Adversarial Training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2107–2116, 2017.

Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse Image Synthesis for Multiple Domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.