Unsupervised Domain Adaptation for Wildlife Image Enhancement: A Task-Aware Approach

CSC2529 Computational Imaging 2025

Authors: K. Hassan Qazi, Aidan Tsang

Abstract

Automated wildlife monitoring relies heavily on deep learning object detectors, yet these models suffer significant performance degradation when processing images captured at night using Near-Infrared (NIR) illumination. Because collecting paired day/night imagery of identical animal poses in the wild is infeasible, supervised learning cannot be applied to bridge this domain gap. This project proposes an unsupervised Image-to-Image (I2I) translation pipeline designed to convert low-quality NIR images into clear, RGB daytime equivalents to improve machine readability.

We introduce a "Task-Aware CycleGAN" that augments the standard Cycle-Consistent Adversarial Network with a novel perceptual loss derived from the intermediate feature maps of a pre-trained MegaDetector (YOLOv5). We compare our approach against standard CycleGAN and state-of-the-art diffusion models (Instruct-Pix2Pix) using the Caltech Camera Traps dataset. While our model successfully hallucinates plausible daytime coloration, quantitative evaluation reveals that the translation process introduces high-frequency artifacts that confuse the detector. Our results indicate that human perceptual quality does not strictly correlate with machine readability, as the baseline NIR images achieved the highest detection accuracy (92.2%) compared to the generated outputs (82.8%–89.1%).

1 Introduction

1.1 Context and Problem

The scale of global biodiversity monitoring has necessitated the use of automated tools, specifically camera traps coupled with deep learning object detectors. While models like MegaDetector have revolutionized data processing, their reliability fluctuates based on environmental conditions. A critical failure mode occurs with images captured at night using Near-Infrared (NIR) flash. These images lack color information and often contain unique noise patterns or blur, causing significant drops in detection accuracy compared to daytime RGB images. The core problem is a domain shift: detectors trained primarily on clear, daytime images struggle to generalize to the NIR domain. Furthermore, this problem cannot be solved via standard supervised image restoration because obtaining “ground truth‘‘ pairs (an animal in the exact same posture in both day and night lighting) is impossible in a wild setting.

1.2 Related Approaches and Limitations

Current solutions generally fall into two categories:

1. General Unsupervised I2I Translation: Frameworks like CycleGAN rely on cycle-consistency to preserve structure without paired data. However, standard CycleGAN optimizes for visual similarity and often hallucinates

artifacts or alters semantic content (e.g., changing a deer's fur pattern), which may confuse detectors.

2. Diffusion Models: State-of-the-art models like Instruct-Pix2Pix offer exceptional textural fidelity. However, they are computationally expensive and prone to "hallucinations'' like inserting background elements or altering animal poses. This compromises the scientific integrity of ecological data.

1.3 Proposed Approach and Contributions

We propose a Task-Aware CycleGAN. Instead of optimizing solely for visual realism, we incorporate a specific feedback signal from the downstream detector. Our method modifies the standard CycleGAN loss function to include a Perceptual Loss calculated between the feature maps of the original and translated images, extracted from the backbone of the MegaDetector itself.

Key Contributions:

1. Task-Aware Architecture: Integration of a frozen object detector (MegaDetector v5a) into the training loop to enforce semantic consistency.
2. Functional Evaluation: A rigorous assessment pipeline that evaluates success based on downstream detection metrics (Accuracy, Precision, Recall) rather than just visual quality (FID).
3. Analysis of Domain Shift: We provide empirical evidence showing that standard enhancement techniques may degrade detection performance due to artifact introduction, challenging the assumption that "daytime-looking" images are inherently better for AI.

2 Related Work

2.1 Low-Light Enhancement and Domain Adaptation

Low-Light Image Enhancement (LLIE) and Unsupervised Domain Adaptation (UDA) are critical for bridging the gap between sensing capabilities and algorithmic interpretation. Traditional LLIE methods often rely on histogram equalization or Retinex theory to adjust illumination [9], while modern deep learning approaches aim to learn mappings from low-light to normal-light distributions. Simultaneously, UDA seeks to minimize the discrepancy between a labeled source domain (e.g., daytime RGB) and an unlabeled target domain (e.g., nighttime NIR) [6]. In the context of wildlife monitoring, the challenge involves not merely enhancing brightness but translating between fundamentally different modalities (single-channel IR to three-channel RGB) without paired supervision.

2.2 Transformer-Based Generators

Recent advancements, such as ITTR [3], replace Convolutional Neural Networks (CNNs) with Vision Transformers. Self-attention mechanisms allow these models to capture global context and long-range dependencies, addressing the locality bias of CNNs. However, despite their performance, Transformers are difficult to deploy in this specific use case. The strict edge computing constraints and limited battery life of remote camera traps necessitate lightweight architectures; the high computational overhead of self-attention mechanisms currently renders them infeasible for in-situ conservation hardware.

2.3 Task-Aware and Diffusion Approaches

In the domain of autonomous driving, Fu et al. (2021) demonstrated that optimizing translation specifically for downstream segmentation tasks yields better results than optimizing for human perception. Conversely, recent generative approaches utilize Diffusion Models, such as Instruct-Pix2Pix [4], which edit images based on text prompts. While visually impressive, our experiments confirm that these models often "hallucinate" new objects, making them unsuitable for scientific data where preserving the exact subject is paramount.

3 Method

3.1 Architecture Overview

Our approach builds upon the CycleGAN framework, consisting of two generators ($G : X \to Y$ and $F : Y \to X$) and two discriminators (D_X and D_Y).

1. Generators: We utilize a U-Net architecture with skip connections. This is chosen over ResNet to better preserve low-level spatial information (edges, outlines) which is critical for object detection.
2. Discriminators: We employ a PatchGAN discriminator, which classifies $N \times N$ patches of an image as real or fake, enforcing high-frequency texture realism.
3. Feature Extractor: To make the model "Task-Aware," we utilize a pre-trained MegaDetector v5a (YOLOv5). This network is frozen during training. We hook into Layer 10 of the backbone to extract intermediate feature maps. Layer 10 was selected because it strikes an optimal balance, containing sufficiently high-level semantic features (e.g., animal shapes) while retaining enough spatial resolution for localization, which is typically lost in deeper layers.

3.2 Loss Function formulation

We optimize a composite loss function that balances adversarial realism, structural consistency, and semantic preservation:

$$L_{total} = \lambda_{cyc} L_{cyc} + \lambda_{percept} L_{percept} + L_{GAN}$$

1. Adversarial Loss (L_{GAN}): Forces the generator to produce images that are indistinguishable from real daytime images.
2. Cycle-Consistency Loss (L_{cyc}): Ensures the translation is reversible, preventing mode collapse.
3. Task-Aware Perceptual Loss ($L_{percept}$): This is our primary technical contribution. We minimize the L1 distance between the MegaDetector feature maps (ϕ_{MD}) of the original image and the output image. This ensures that the features used by the detector to identify animals are preserved through the translation cycle.

where

$$L_{percept} = \|\phi_{MD}(x) - \phi_{MD}(G(x))\|_1$$

3.3 Implementation Details

Our code is available at: https://github.com/KhuzemahHQ/Wildlife_Domain_Adapt

The model was implemented in PyTorch. We used the Caltech Camera Traps (CCT) dataset [5], specifically the Missouri subset, splitting it into Domain A (Night/NIR) and Domain B (Day/RGB). Images were resized to 512×512. We trained the model for 10 epochs using the Adam optimizer ($lr = 0.0002$, $\beta_1 = 0.5$). It should be noted that training was halted at 10 epochs due to the time and computational constraints associated with the course timeline, rather than this being the experimentally determined point of convergence.

Architecture of the Task-Aware CycleGAN

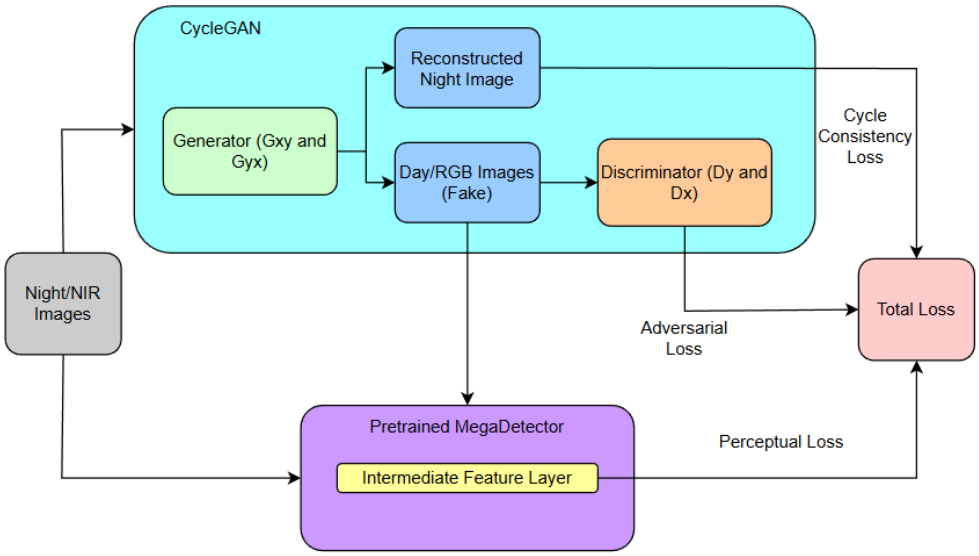

Figure 1

4 Results

4.1 Evaluation Methodology

Since no ground truth exists for unpaired translation, we rely on Functional Utility. We run the enhanced images through MegaDetector v5a and compare detection performance against ground truth bounding box annotations provided in the CCT dataset. We report Accuracy, Precision, Recall, and F1 Score on whether animals are detected at a confidence threshold of 0.25.

4.2 Baselines

We compare our method against two baselines:

1. Base NIR: The original, unenhanced nighttime images (establishing the lower bound).
2. Standard CycleGAN: Vanilla CycleGAN without the perceptual loss.

4.3 Qualitative Results

Visually, the domain adaptation succeeds in colorizing NIR images.

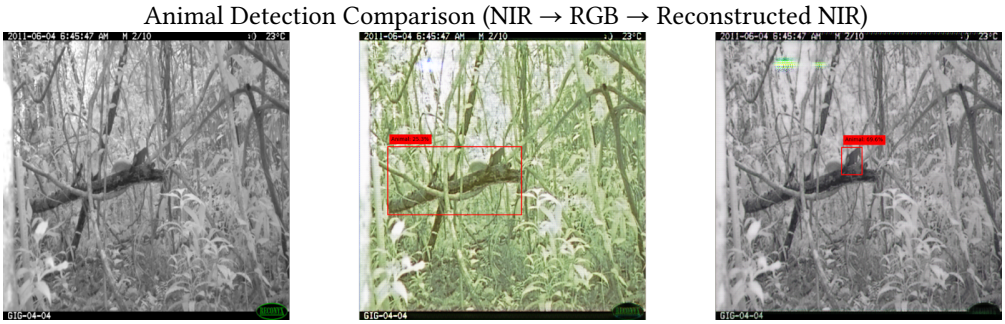

Figure 2

Our CycleGAN successfully mimics the daytime color palette but introduces high-frequency noise/artifacts in uniform regions.

4.4 Quantitative Results

The quantitative results on the downstream detection task are summarized in the table. Counter-intuitively, the Base NIR images performed best (92.2% Accuracy). While our model successfully translated the domain visually, the detection accuracy dropped to 82.8% - 89.1%. This leads to critical insights regarding the failure modes of generative enhancement:

1. Noise Amplification and Artifacts: The GAN generation process introduces high-frequency adversarial noise and "checkerboard artifacts" characteristic of deconvolutional layers. While these textures often appear plausible to the human eye, they act as adversarial attacks on the detector's convolutional filters, disrupting feature extraction.
2. Insufficiency of Standard Image Processing: It is worth noting that simpler baselines, such as Histogram Equalization, were considered but deemed insufficient. Such methods operate only on pixel intensity and cannot hallucinate the chromatic information required to bridge the domain shift from single-channel NIR to three-channel RGB, necessitating the use of generative approaches.

Table 1: *Performance Metrics Comparison across Image Types*

Image Type	Accuracy	Precision	Recall	F1 Score
Base NIR (Original)	**0.922**	0.978	**0.917**	**0.946**
CycleGAN Generated Daytime (No Perceptual)	0.891	**1.000**	0.854	0.921
CycleGAN Reconstructed Nighttime (No Perceptual)	0.875	0.976	0.854	0.911
CycleGAN Generated Daytime (Perceptual $\lambda = 5$)	0.828	**1.000**	0.771	0.871
CycleGAN Reconstructed Nighttime (Perceptual $\lambda = 5$)	0.797	0.927	0.792	0.854
CycleGAN Generated Daytime (Perceptual $\lambda = 10$)	0.844	0.975	0.812	0.886
CycleGAN Reconstructed Nighttime (Perceptual $\lambda = 10$)	0.906	**1.000**	0.875	0.933

5 Discussion

5.1 Divergence of Perception vs. Readability:

This research highlights a crucial divergence in computer vision: Perceptual Quality ≠ Machine Readability. While we successfully enhanced images to look pleasing to humans (coloring night images), this "enhancement" degraded machine performance. This implies that for conservation AI, simpler, noise-free images are preferable to synthetic, colored images that contain generative artifacts.

This core quantitative finding is a critical challenge to the conventional wisdom that "daytime-looking" images are inherently superior for machine vision. Our Task-Aware CycleGAN was successful in achieving visual realism and semantic consistency (as enforced by the $L_{percept}$ loss), yet the translation process introduced high-frequency noise and artifacts. This suggests that the MegaDetector, which was primarily trained on RGB, may have learned robust filters for common NIR degradations, and the generated artifacts from the I2I translation are a novel form of noise that the model struggles to generalize over.

5.2 Model Convergence & Training:

Given that the project's original 6-week time constraint is no longer a factor, a crucial next step for our Task-Aware CycleGAN is to continue training beyond the initial 10 epochs. The abrupt cessation of training may have prevented the model from fully converging and resolving the subtle artifacts that degraded detector performance. Training for an additional 20-30 epochs would provide the generators and discriminators the necessary time to reach a more stable Nash equilibrium, potentially smoothing out the high-frequency noise while better preserving semantic details, which could finally translate to an increase in the downstream detection metrics.

5.3 The Role of Diffusion Models:

In our related work analysis, we considered state-of-the-art Diffusion Models (specifically Instruct-Pix2Pix) for image translation. While the model was exceptional at generating high-frequency detail and achieving visually impressive super-resolution in preliminary tests, we made a strategic decision not to pursue this direction for the main project. The primary reasons were their computational demands: they took significantly longer to train and run inference compared to the lightweight CycleGAN framework.

Despite this, the potential of diffusion models for restoring and enhancing medical or ecological imagery is undeniable. Future work will prioritize exploring how a smaller, more dedicated diffusion model. A fine-tuned tool specifically for restoration, rather than general editing, could be deployed. This could include a conditional diffusion process guided by NIR input, potentially offering the best balance of detail preservation and domain translation without the prohibitive computational cost of larger, multimodal models.

5.4 Real-World Impact:

Beyond the academic exercise, this research is driven by real-world conservation needs. To bridge the gap between research and practice, we have identified a prospective avenue for deployment with conservationists in Nepal. As they actively seek resource-efficient solutions for remote camera traps, their operational needs closely mirror the design philosophy of our task-aware model. We intend to investigate the feasibility of applying our framework to their data, which would serve as a robust test of our model's capabilities in a high-impact, resource-constrained

environment.

5.5 Limitations

- The primary limitation is the trade-off between the Adversarial Loss (which introduces texture/noise to mimic "daytime") and the Perceptual Loss (which tries to keep features clean). Currently, the Adversarial noise overpowers the structural preservation. Additionally, our dataset size was constrained for the course timeline, potentially limiting the GAN's ability to generalize.
- Adversarial Artifacts vs. Machine Readability: A primary limitation is the inherent conflict within our composite loss function. While the Adversarial Loss (L_{GAN}) successfully pushes the generator to match the texture and color distribution of daytime images, it introduces high-frequency artifacts and stochastic noise. While these artifacts are often perceptually negligible to human observers, they act as adversarial perturbations to the convolutional filters of the downstream detector, degrading detection accuracy despite visual improvement.
- Dependency on Pre-trained Feature Extractor: Our task-aware approach is inextricably linked to the biases and architecture of the frozen MegaDetector v5a. We assume that the intermediate feature maps (Layer 10) contain the optimal semantic representation for all wildlife. However, because MegaDetector was trained primarily on RGB images, its internal feature representations may not be fully invariant to the specific artifacts generated during NIR-to-RGB translation, effectively capping our model's performance ceiling based on the detector's own robustness.
- Lack of Ground Truth Metrics: Due to the physical impossibility of capturing paired wild animal data, we rely heavily on proxy metrics. We cannot utilize standard full-reference metrics like PSNR or SSIM6. While our functional evaluation is pragmatic, it is an indirect measure of image quality. We cannot definitively prove that the reconstructed biological details (such as fur patterns) are faithful to the specific individual animal, posing a risk for tasks requiring individual re-identification.

5.6 Future Work

- Dynamic Loss Weighting: Implementing a schedule where $\lambda_{percept}$ increases over time to prioritize structure once the color mapping is learned.
- Offline Data Augmentation: Rather than using the GAN for inference (which adds computational cost), the most promising path is to use the model offline to generate a massive synthetic dataset of "fake day" images. This synthetic data could then be used to fine-tune a new version of MegaDetector. By training the detector on the artifacts, it would learn to be robust to them, allowing for improved performance without altering the deployment pipeline.
- Background Replacement via Semantic Compositing: Leverage the stationary nature of camera traps to decouple the subject from the environment by maintaining a "bank" of daytime reference images for each location; using a segmentation model to composite the animal onto a real daytime background would guarantee photorealistic surroundings without generative hallucinations, allowing the pipeline to focus exclusively on local contrast enhancement of the animal itself.

Bibliography

1. Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV).
2. Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal Unsupervised Image-to-Image Translation (MUNIT). Proceedings of the European Conference on Computer Vision (ECCV).
3. Li, B., et al. (2022). ITTR: Unpaired Image-to-Image Translation with Transformers. ArXiv.
4. Brooks, T., Holynski, A., & Efros, A. A. (2023). Instruct-Pix2Pix: Learning to Follow Image Editing Instructions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
5. Beery, S., Morris, D., & Yang, S. (2019). The Caltech Camera Traps Dataset. ECCV Workshops.
6. Ganin, Y., & Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. International Conference on Machine Learning (ICML).
7. Hoffman, J., et al. (2018). CyCADA: Cycle-Consistent Adversarial Domain Adaptation. International Conference on Machine Learning (ICML).
8. Liu, M. Y., & Tuzel, O. (2016). Coupled Generative Adversarial Networks. Advances in Neural Information Processing Systems (NeurIPS).
9. Lore, K. G., Akintayo, A., & Sarkar, S. (2017). LLNet: A Deep Autoencoder Approach to Natural Low-light Image Enhancement. Pattern Recognition.
10. Wei, C., et al. (2018). Deep Retinex Decomposition for Low-Light Enhancement. British Machine Vision Conference (BMVC).
11. Guo, C., et al. (2020). Zero-DCE: Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
12. Jiang, Y., et al. (2021). EnlightenGAN: Deep Light Enhancement without Paired Supervision. IEEE Transactions on Image Processing.