frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 28 August 2026
DOI 10.3389/fevo.2026.1940923

From noise to nature: diffusion models to advance wildlife detection in aerial imagery

Henry Sun[1*], Holly R. Houliston[2,3], Jiayi Zhou[4], Max Niu[1], Brinnae Bent[4] and David W. Johnston[1]

[1]Division of Marine Science and Conservation, Nicholas School of the Environment, Duke University Marine Laboratory, Beaufort, NC, United States, [2]British Antarctic Survey, Cambridge, United Kingdom, [3]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom, [4]Pratt School of Engineering, Duke University, Durham, NC, United States

Check for updates

OPEN ACCESS

EDITED BY
Nataša Popović,
Institute for Biological Research "Siniša Stanković", Serbia

REVIEWED BY
Umar Yahya,
Universiti Islam Sultan Sharif Ali, Brunei
Haochen Sun,
Zhengzhou University, China

*CORRESPONDENCE
Henry Sun
henry.sun@duke.edu

RECEIVED 17 July 2026
REVISED 08 August 2026
ACCEPTED 14 August 2026
PUBLISHED 28 August 2026

CITATION
Sun H, Houliston HR, Zhou J, Niu M, Bent B and Johnston DW (2026) From noise to nature: diffusion models to advance wildlife detection in aerial imagery.
Front. Ecol. Evol. 14:1940923.
doi: 10.3389/fevo.2026.1940923

Automated wildlife detection in aerial imagery can expand the scale and efficiency of ecological monitoring, but model development is often constrained by limited training data, uneven class representation, and poor coverage of diverse environmental conditions. These constraints are especially acute for rare, elusive, or endangered species, where large, diverse, and well-annotated image datasets may be difficult or impossible to obtain. Synthetic imagery generated by fine-tuned diffusion models could help address these limitations, but its utility for wildlife detection remains insufficiently tested. Here, we evaluated whether diffusion-generated imagery can improve object detection performance in a remote sensing workflow using humpback whales (*Megaptera novaeangliae*) in drone imagery as a benchmark case study to simulate data limitations relevant to rare or poorly observed species. We fine-tuned Stable Diffusion XL 1.0 using low-rank adaptation on a small independent set of real whale images to generate abundant synthetic aerial imagery, and trained YOLOv8 object detection models using controlled combinations of real and synthetic data. We found that the addition of synthetic imagery improved localization accuracy measured using mAP50–95 when real training data were limited in quantity or diversity, provided little additional benefit when real data were sufficient, and substantially reduced all performance metrics when used alone. These findings suggest that diffusion-generated synthetic imagery can strengthen detection models under data-limited conditions but should complement, rather than replace, real imagery. We conclude by outlining practical applications and key research priorities for integrating synthetic data into wildlife detection workflows.

KEYWORDS
diffusion, drone, generative AI, image augmentation, machine learning

1 Introduction

For ecologists, identifying animals in imagery is crucial to assess biodiversity, monitor population trends, and document behaviors at large scales, often informing evidence-based conservation and management (Michener and Jones, 2012; Johnston, 2019). Automated methods for animal identification in imagery using machine learning (ML) reduce the cost and logistical overhead associated with manual annotation (Seymour et al., 2017; Gray et al., 2018; Weinstein, 2018). ML detection models for animal recognition require large image training datasets that capture the range of variation present within real-world scenarios (Deep and Dash, 2019; Guirado et al., 2019). However, procuring a vast quantity of representative training images can be difficult or near-impossible for certain tasks (Shorten and Khoshgoftaar, 2019). Developing robust ML wildlife detection models is thus often constrained by limited and imbalanced datasets, particularly for rare, elusive, or low-detectability species, where scarce or uncertain annotations increase the risk of overfitting and poor generalization (Langenkämper et al., 2019, Leevy et al., 2018, Delisle et al., 2022). Paradoxically, the automated detection models needed to help monitor and conserve rare and endangered species require large, diverse image datasets for robust training and reliable performance, yet the scarcity and limited observability of these species often make it difficult to obtain enough representative imagery.

ML researchers commonly address data deficiency problems like this through dataset augmentation, where additional images for training detection models are generated through manipulations of existing images (Mumuni and Mumuni, 2022). Conventional dataset augmentation is widely used to increase the size and variation within training datasets through transformations such as color channel shifts, rotations, flips, or crops, thereby improving the performance of ML detection models (Paszke et al., 2019; Shorten and Khoshgoftaar, 2019). Alternatively, synthetic dataset augmentation produces training images *de novo* and can offer fine-grained control and customizability of real-world conditions (Beery et al., 2020). In particular, generative artificial intelligence (AI) models can learn complex spatial and spectral distributions from training data to produce highly realistic synthetic imagery, with potential for rapid and accurate synthetic dataset augmentation (Liu et al., 2024; Rafiq et al., 2025).

Diffusion models have emerged as the eminent class of AI image generation models, outperforming architectures such as generative adversarial networks (GANs) and variational autoencoders (VAEs) in image quality and diversity (Dhariwal and Nichol, 2021; Muhammad et al., 2023; Müller-Franzes et al., 2023). Diffusion models progressively add noise to training images and then learn to iteratively reverse the noising process to generate novel samples (Ho et al., 2020; Croitoru et al., 2023; Rombach et al., 2022). In contrast, GANs, which have historically dominated synthetic image generation in ecology (Goodfellow et al., 2014; Lu et al., 2019), rely on an adversarial training process between a generator that produces synthetic samples and a discriminator that distinguishes real images from generated ones. While GANs can produce sharp and visually convincing outputs, they are prone to training instabilities and mode collapse, where the generator produces a limited range of samples that lack meaningful diversity (Jabbar et al., 2022). Alternatively, VAEs encode data into a latent distribution and generate images by sampling and decoding from this space (Kingma and Welling, 2022), but their reliance on normally distributed latent variables often leads to blurred outputs due to averaging distinct features (El-Kaddoury et al., 2019). Many diffusion models use a VAE to compress images into a lower-dimensional space, but learn to generate images by reversing a gradual noising process within that space, rather than relying on a VAE alone to decode samples from a constrained latent distribution (Podell et al., 2023).

Large foundation diffusion models such as OpenAI's DALL-E or Stability AI's Stable Diffusion have been pretrained on massive image–text datasets, enabling broad image generation from text prompts. While zero-shot image generation can suffice for certain applications (Durand et al., 2026), for domain-specific visually nuanced tasks, higher-quality synthetic images can be generated using model fine-tuning. Fine-tuning refers to adapting a pre-trained model to a specific task or concept using a smaller, targeted dataset, enabling the generation of specialized images while retaining general knowledge encoded in the base model (Ding et al., 2023; Ruiz et al., 2023). Low-Rank Adaptation (LoRA) is a fine-tuning method that adapts large foundation models to new tasks without updating all parameters by freezing the original weights and learning a small set of low-rank "adapter" matrices that encode task-specific adjustments (Hu et al., 2021). In text-to-image diffusion models, these adapters are typically inserted into layers that link textual prompts to visual features, enabling the model to learn new styles or subjects with a fraction of the compute and storage costs of full fine-tuning (Cuenca and Paul, 2023).

Synthetic imagery has previously been used to augment wildlife detection models through several distinct strategies. Gaur et al. (2023) generated synthetic whale imagery using 3D simulation, Durand et al. (2026) used zero-shot text-to-image prompting with the generative model DALL-E 2 for aerial muskox detection, and Duporge et al. (2026) trained a generative diffusion model *de novo* for satellite-based marine fauna detection. In contrast, we evaluate LoRA fine-tuning of a pretrained foundation diffusion model as a computationally efficient approach for generating domain-specific wildlife imagery from a small reference dataset. We further extend these previous studies by systematically benchmarking the downstream utility of synthetic imagery across controlled experimental scenarios that manipulate the quantity and diversity of available real training data.

The present study uses humpback whale (*Megaptera novaeangliae*) detection in drone imagery as a benchmark system to evaluate this approach under simulated data constraints. Although humpback whales are comparatively abundant and well represented in aerial imagery datasets, we deliberately fine-tuned Stable Diffusion XL 1.0 using a small, independently sourced set of real humpback whale images and used the resulting model to generate synthetic imagery for detector training. We then trained YOLOv8 models using controlled combinations of real and synthetic imagery, including scenarios with reduced real data quantity and flight-level diversity. This experimental design allowed us to test whether synthetic imagery could supplement or replace real training data and whether its utility depended on the availability and diversity of real imagery. All models were evaluated against the same independent real image validation and test sets, enabling direct comparison of when synthetic imagery provided measurable benefits, offered little additional value, or incurred performance costs.

2 Materials and methods

2.1 Overview

We evaluated whether diffusion-generated imagery could improve humpback whale detection in drone imagery by generating synthetic training images, combining them with controlled subsets of real imagery, and comparing YOLOv8 model performance across nine training scenarios. The workflow comprised four steps: (1) synthetic image generation and annotation, (2) real image preparation and dataset comparison, (3) construction of experimental training and testing sets, and (4) model training and evaluation. The full experimental design is summarized in Figure 1 and Table 1.

Briefly, a small independent set of real humpback whale images (n = 27, indicative of a very small population of animals) was used to fine-tune a diffusion model and produce abundant synthetic humpback whale images for model training (Figure 1, step 1). These synthetic images were then combined with real drone imagery in different configurations (Figure 1, step 2) to create a set of controlled training scenarios (Figure 1, E1–E9). A diverse real-only model served as the baseline (E1; Figure 1). The remaining experiments tested whether model performance changed when standard augmentation was removed (E2), when synthetic images replaced some or all real images (E3–E4), when synthetic images were added to an already strong real dataset (E5), and when synthetic images were added to restricted real datasets that simulated rare species conditions with fewer flights, fewer images, or both (E6–E9). All models were trained using the same YOLOv8 framework and evaluated on the same independent real image validation and test sets (Figure 1, steps 3-4), allowing performance differences to be attributed to training data composition rather than changes in evaluation data.

2.2 Synthetic image generation and annotation

We generated 2,112 synthetic humpback whale images at 1024 × 1024 pixels using a fine-tuned pre-trained foundation diffusion model, Stable Diffusion XL 1.0 (Figure 2). This open-weight diffusion model has a larger parameter count and more refined training strategy compared to earlier Stable Diffusion models, enabling the generation of high-fidelity images from text prompts (Podell et al., 2023). For LoRA fine-tuning, 27 real drone images of humpback whales were selected from a larger dataset comprising 602 images of 228 individuals collected along the Western Antarctic Peninsula between February 24th and March 11th, 2019 (Bierlich et al., 2022). Briefly, these images were collected using a FreeFly Alta 6 multirotor drone fitted with a Sony Alpha a5100 camera with an APS-C (23.5 × 15.6 mm) sensor, 6,000 × 4,000 pixel resolution, and a 35 mm lens. These images were independent from the real images used for detection model training described below and were deliberately chosen to span variation in animal orientation, weather, and lighting conditions. Descriptive text captions summarizing image content were generated for each real image used in fine-tuning.

LoRA fine-tuning of Stable Diffusion XL 1.0 was done using the train_dreambooth_lora_sdxl.py script provided by the open-source

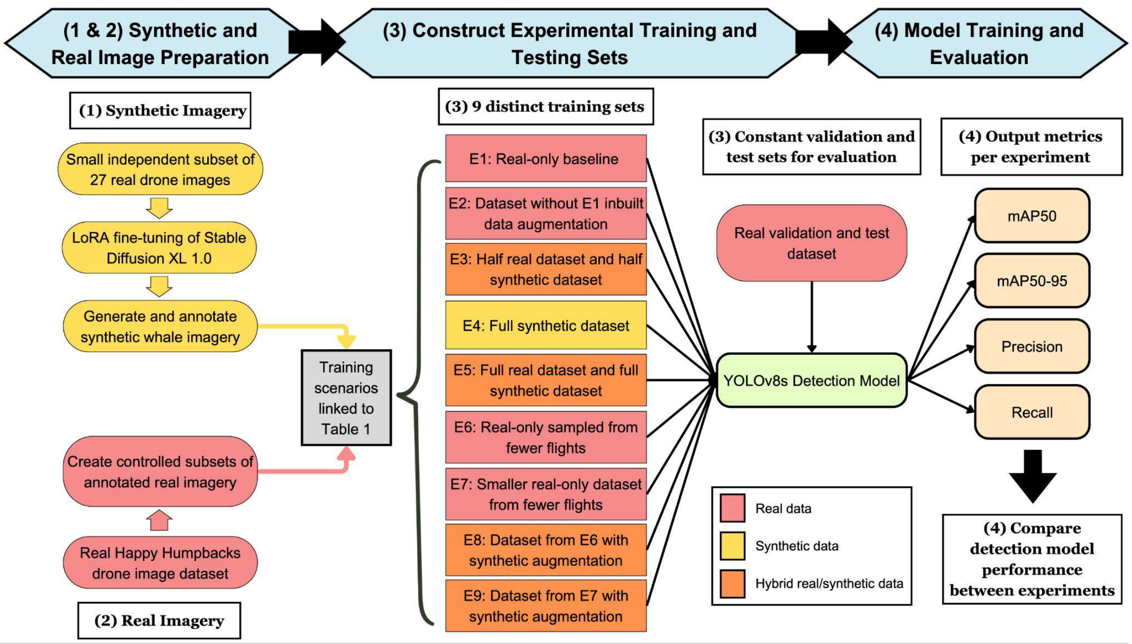

FIGURE 1
Experimental design used to test synthetic data augmentation for wildlife detection in this study. A small independent set of real humpback whale drone images was used to fine-tune a diffusion model and generate synthetic training images (1). In parallel, the real humpback whale drone image dataset was subset into controlled training datasets (2). These synthetic and real datasets were then used to create nine separate training datasets [E1–E9], corresponding to the nine experiments listed in Table 1 (3). All models were evaluated using four metrics on the same independent real validation and test sets, so differences in performance reflect training data composition rather than differences in evaluation data (4).

TABLE 1 **Explanation of the nine YOLOv8 training experiments conducted.**

Experiment ID	Training setup	Primary comparison	Purpose
E1	Diverse real-only dataset with standard YOLOv8 augmentation	Reference experiment	Establishes the real-data baseline for all comparisons.
E2	Same real dataset as E1, but with built-in YOLOv8 data augmentation disabled	E2 vs. E1	Tests the contribution of standard YOLOv8 data augmentation approaches.
E3	Half real data + half synthetic data, with total training size matched to E1	E3 vs. E1	Tests whether synthetic images can partially replace real images while holding total training size constant.
E4	Synthetic-only training data	E4 vs. E1	Tests whether synthetic images can fully replace real training imagery.
E5	Full real baseline dataset plus the full synthetic dataset	E5 vs. E1	Tests whether adding synthetic images improves performance when real data are already diverse and sufficient.
E6	Same number of real whale instances as E1, but drawn from fewer flights	E6 vs. E1	Tests the effect of reduced real-data diversity while holding instance count similar.
E7	Smaller real dataset drawn from fewer flights	E7 vs. E1	Tests the combined effect of reduced real-data size and reduced flight diversity.
E8	Reduced-diversity real dataset plus the full synthetic dataset	E8 vs. E6 and E1	Tests whether synthetic images improve models trained on limited-diversity real data.
E9	Smaller reduced-diversity real dataset plus a matched synthetic subset	E9 vs. E7 and E1	Tests whether synthetic images improve models trained under both reduced size and reduced diversity.

Experiment IDs correspond to the labeled training scenarios in Figure 1. The design tests three common data limitations encountered in rare species detection: limited real image quantity, limited survey diversity, and uncertainty about whether synthetic imagery should supplement or replace real data. Exact counts of the number of instances, images, and flights for each training set as well as for the validation and test sets are provided in Supplementary Table 1.

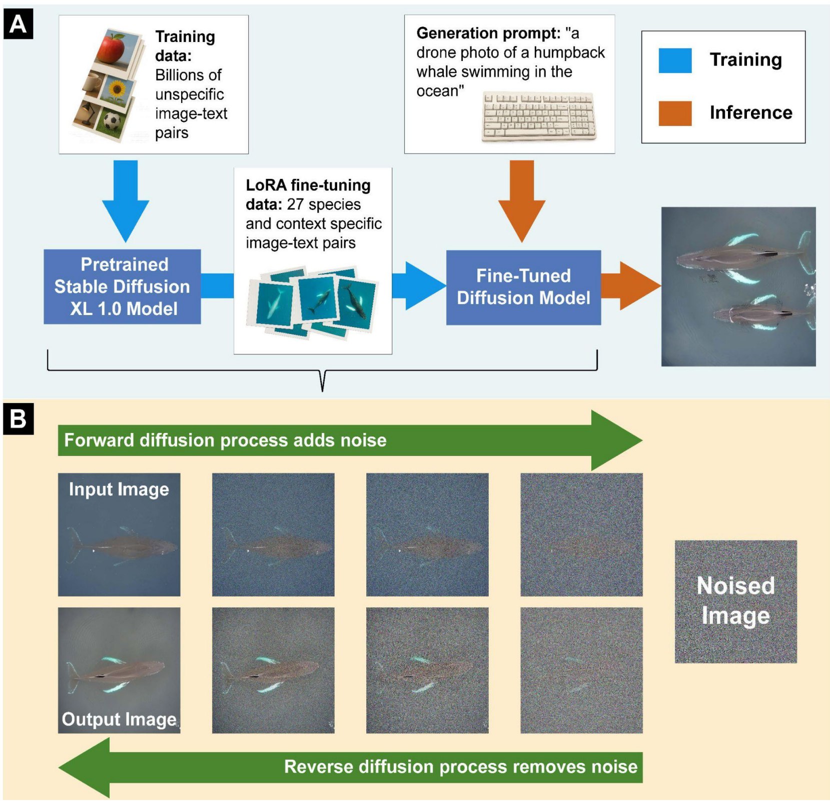

FIGURE 2
(A) Process for generating synthetic aerial imagery of humpback whales using a fine-tuned Stable Diffusion XL 1.0 model. Blue arrows denote steps where model weights are adjusted in training, and orange arrows denote steps where outputs are generated using inference of trained models. **(B)** Schematic representation of the diffusion process, including the forward step which adds noise to an image, and the reverse step where the model learns during training to reconstruct an image by iteratively denoising random noise. The model can then generate novel images from noise that are consistent with the training data distribution.

Diffusers library v0.34.0 (von Platen et al., 2022). Fine-tuning used a batch size of 2, gradient accumulation of 1, Prodigy optimization, a constant learning rate scheduler with no warmup, and 1,350 training steps. The LoRA rank was set to 16, a commonly used value for diffusion model fine-tuning that provides sufficient adaptation capacity while limiting overfitting. Gradient clipping was applied with a maximum norm of 1.0, and training used bfloat16 mixed precision, cached latents, and gradient checkpointing. Textual inversion of the text encoder was enabled for 50% of the training process. The instance prompt "drone image of a <s0><s1> in the ocean, clear water, visible pectoral fins, ultra-realistic" was used. The tokens <s0><s1> are learnable identifiers that acquire the visual characteristics of the fine-tuning images during training, allowing the model to generate new images of humpback whales while preserving the descriptive context provided by the remainder of the prompt. The fine-tuned LoRA model was then used to generate synthetic images with the text prompt "a drone photo of a humpback whale swimming in the ocean." Fine-tuning Stable Diffusion XL 1.0 required approximately 30 minutes, and generation of the 2,112 synthetic images required approximately 1 hour, both using a single NVIDIA A100 GPU runtime.

Synthetic images were annotated by two experienced reviewers using LabelMe 5.11.1 (Wada, 2026), with bounding boxes placed around the visible extent of each whale. The synthetic image dataset was divided between two annotators, and each annotation set was independently cross-reviewed by the other annotator to assess accuracy and consistency before the final annotations were confirmed. When hallucinated features such as duplicated body parts or anatomically unrealistic structures were present (as in Supplementary Figure 1A), separate bounding boxes were drawn around each visually distinct whale-like structure. This approach was adopted to prevent the detection model from learning incorrect object extents and was intended solely to constrain object localization during training.

2.3 Real image preparation and dataset comparison

Real imagery was sampled from the *Happy Humpbacks* dataset, a labeled and annotated multi-rotor drone dataset collected at Palmer Station along the Western Antarctic Peninsula (Houliston et al., 2026). The dataset includes imagery from 55 drone flights and comprises 5,281 curated images containing 10,401 annotated humpback whale instances. The *Happy Humpbacks* images were taken with a DJI Phantom 4 Pro multirotor drone equipped with a default gimbal-stabilized RGB camera (1" CMOS sensor; 5472 x 3648 pixels) during January-March 2020 (Houliston et al., 2026). Prior to detection model training, all real images were resized to 1024 × 1024 pixels, with padding applied to preserve the original aspect ratio.

Because synthetic images may differ systematically in background appearance, we compared background color distributions between real images used for detection model training, real images used for LoRA fine-tuning, and synthetic images. Specifically, mean red, green, and blue (RGB) pixel values and lightness, green-red, and blue-yellow values in CIELAB color space (LAB) were calculated for background regions excluding bounding boxes within each image using OpenCV 4.13.0.92. We assessed differences in RGB and LAB distributions among (1) a random sample of 2,112 real training images from the *Happy Humpbacks* dataset, (2) the 27 real images used for LoRA fine-tuning, and (3) the full dataset of 2,112 synthetic images using two-sample Kolmogorov-Smirnov tests for each color channel (R, G, B, and L, A, B). We also used principal component analysis (PCA) in scikit-learn 1.8.0 to visualize relationships among image-background color distributions across datasets.

2.4 Construction of experimental training and testing sets

We generated training data subsets for the nine linked experiments summarized in Table 1. Because our objective was object detection rather than image classification, both synthetic and real images could contain multiple whales, with each whale having its own manually annotated bounding box (hereafter referred to as an instance), resulting in a greater number of instances than images. The exact image, instance, and flight counts for each experiment are provided in Supplementary Table 1. The training sets for Experiments E1, E2, and E5, and the validation and test sets, were generated by sampling subsets from the *Happy Humpbacks* dataset. To do this, images in the dataset were first grouped by flight ID, and independent sampling was performed within each flight using a random starting point followed by regular interval (stride) sampling. The stride length was adjusted according to the number of images in each flight so that all flights remained proportionally represented in the final subset. A subset of these sampled images was then selected to preserve the distribution of whale instances per image seen in the original *Happy Humpbacks* dataset. This approach reduces autocorrelation among consecutive drone frames in the training datasets while maintaining representation across flights.

To simulate data scarcity and reduced survey diversity for Experiments E6–E9, real training subsets were generated with a target number of instances while minimizing the number of flights included. Images were grouped by flight ID, and candidate flight groups were sampled iteratively, starting with the smallest number of flights. For each candidate set, images were searched to identify a subset whose total instance count matched the target instance count. The subset with the fewest flights meeting the target instance count was selected. In Experiments E4, E5, and E8, the full synthetic dataset of 2,112 images was used.

For Experiments E3, E7, and E9, datasets were partitioned into two subsets, A and B, using stratified random sampling based on instance-per-image counts. Images were grouped by instance count, randomly shuffled within each group, and split using a 50:50 ratio. A subsequent correction step involving the addition, removal, or swapping of images was applied to better match target instance totals while preserving the instance-per-image distribution. Detection models were trained independently on the A and B subsets to evaluate sensitivity to training data sampling. Because differences in detection model performance between paired training subsets were much smaller than the treatment effects

(Supplementary Figure 2), results for these experiments are reported as the average across A and B subsets to simplify interpretation.

2.5 Model training and evaluation

We trained YOLOv8s detection models for all nine experiments (Redmon et al., 2016). YOLO architectures are widely used for wildlife detection tasks (Dave et al., 2023; Roy et al., 2023; Chen et al., 2024), and we selected YOLOv8s for its balance among accuracy, inference speed, ease of deployment in real-world pipelines, and suitability for our dataset size and detection task. Across all experiments, we used a shared baseline configuration of 100 training epochs and a batch size of 32. We froze the first 21 layers to fine-tune only the detection head for more stable optimization. All other hyperparameters were kept at the default YOLOv8s settings. All training runs were repeated across six random seeds, 0–5, for each experiment. For Experiment E2, all default YOLOv8 data augmentation was disabled by setting augmentation-related parameters to zero or false. Examples of built-in YOLOv8 augmentation include flips, color contrast shifts, and blurs. All experiments were executed on a 64-bit system using an NVIDIA GeForce RTX 4090 GPU with 24 GB VRAM, with the full set of experiments requiring an average of 6.76 GPU hours per random seed. Training and validation loss curves for all experiments are provided in Supplementary Figure 3.

Model performance was evaluated using mean average precision at a 50% intersection-over-union threshold (mAP50), mean average precision across thresholds from 0.50 to 0.95 intersection-over-union (mAP50–95), precision, and recall. mAP50 quantifies detection accuracy under a single moderate overlap criterion, where predicted bounding boxes are considered correct if their intersection-over-union with matching ground-truth boxes is at least 50%. In contrast, mAP50–95 averages precision across multiple intersection-over-union thresholds from 0.50 to 0.95 in increments of 0.05, providing a more stringent assessment of localization accuracy. Precision was calculated as the proportion of true positives relative to all predictions, reflecting the tendency to avoid false positives. Recall was calculated as the proportion of true positives relative to all ground-truth instances, reflecting the ability to minimize missed detections. Following confirmation of normality with Shapiro-Wilk tests and homogeneity of variance with Levene's tests, we compared differences in detection model performance for all four metrics for each experimental comparison described within Table 1 using Student's t-tests at a significance level of $\alpha = 0.05$. P-values were adjusted for multiple tests within each experimental comparison using Holm's correction. We also report the F1-Score for each model, defined as (2 * *Precision* * *Recall*)/(*Precision* + *Recall*), but did not analyze these results further given the near-identical overlap with trends in precision and recall. All statistical analysis was performed in R 4.4.3 (R Core Team, 2025).

3 Results

3.1 Dataset color channel comparisons

Significant differences were observed between real images (Figures 3A, B), LoRA training images (Figures 3C, D), and synthetic images (Figures 3E, F) for all LAB channels (two-sample K–S tests; all $p < 0.001$, Supplementary Table 2, Supplementary Figure 4). Similarly, significant differences were observed across all pairwise dataset comparisons for all three RGB channels (all $p < 0.001$; Supplementary Table 2, Supplementary Figure 5). PCA revealed that synthetic images formed a distinct, tightly clustered group, reflecting lower variation than either the real or LoRA fine-tuning datasets (Figure 3G). Loading plot values for PCA showed that PC1 was dominated by the L channel (loading = 0.997), with minimal

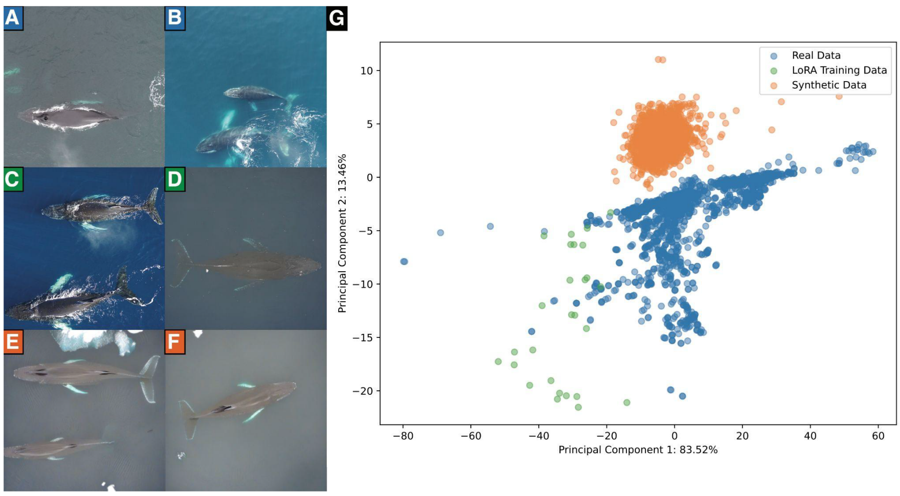

FIGURE 3
Differences in imagery and image color characteristics between the real dataset, LoRA training dataset, and the synthetic dataset used in this study. **(A, B)** Example images from the real training dataset. **(C, D)** Example images from the LoRA training dataset. **(E, F)** Example images from the synthetic training dataset. Images were cropped to 512x512 pixels for this figure. **(G)** PCA of LAB channels from LAB color space for real, LoRA training, and synthetic images.

contribution from the A and B channels. PC2 had contributions from A (loading = 0.766) and B (loading = 0.639) channels.

3.2 Overall model performance with synthetic imagery

We trained detection models on a total of nine different training datasets to evaluate the influence of synthetic imagery and data augmentation on detection model performance. Results showing mean performance across all metrics for the experiments are available in Table 2.

The baseline model trained on 2,693 real instances achieved an mAP50 of 0.7770, mAP50–95 of 0.5008, precision of 0.8744, and recall of 0.7005. Disabling in-built data augmentation approaches using the same training dataset in Experiment 2 substantially reduced detection performance (Supplementary Figure 6). Specifically, mAP50 decreased from 0.7770 to 0.5817 (p-adj. = 1.76×10^{-4}) and mAP50–95 from 0.5008 to 0.2778 (p-adj. = 1.76×10^{-4}). Precision and recall also declined significantly, from 0.8744 to 0.6829 (p-adj. = 0.0014) and from 0.7005 to 0.5321 (p-adj. = 1.34×10^{-6}), respectively.

The impact of replacing 50% of real training instances with synthetic data in Experiment 3 resulted in reductions in mAP50 (0.7770 vs. 0.7616, p-adj. = 0.0043) and mAP50–95 (0.5008 vs. 0.4867, p-adj. = 2.52×10^{-4}) (Figure 4). In contrast, precision (p-adj. = 0.23) and recall (p-adj. = 0.21) did not differ significantly (Supplementary Figure 7). Replacing real training data entirely with synthetic data in Experiment 4 decreased performance substantially across all metrics, representing the lowest overall performance among all models tested in this study (Figure 4). mAP50 decreased to 0.4070 (p-adj. = 1.88×10^{-8}) and mAP50–95 to 0.1896 (p-adj. = 1.88×10^{-8}). Precision and recall decreased to 0.5838 (p-adj. = 1.88×10^{-8}) and 0.3779 (p-adj. = 3.4×10^{-9}), respectively.

Doubling the training dataset size by adding an equal number of synthetic instances in Experiment 5 had no effect on model performance. While this configuration yielded the highest mean values for mAP50 (0.7782), mAP50–95 (0.5014), and recall (0.7060) across all experiments, these increases were not statistically significant relative to the baseline (mAP50: p-adj. = 1; mAP50–95: p-adj. = 1; precision: p-adj. = 1; recall: p-adj. = 1) (Supplementary Figure 8).

3.3 Model performance under reduced dataset diversity

Reducing the diversity of training data in Experiment 6 caused significant declines in mAP50 (0.7464, p-adj. = 0.0011) and mAP50–95 (0.4542, p-adj. = 1.17×10^{-5}) (Figure 5). Precision and recall decreased to 0.8460 (p-adj. = 0.11) and 0.6888 (p-adj. = 0.17), although these changes were not significant (Supplementary Figure 9). Augmenting this reduced-diversity dataset with synthetic data in Experiment 8 increased mAP50 to 0.7557 (p-adj. = 0.39) but not significantly, while mAP50–95 significantly increased to 0.4698 (p-adj. = 0.0024). Precision and recall also increased slightly, to 0.8543 (p-adj. = 0.62) and 0.6909 (p-adj. = 0.82), respectively, but these differences were not significant.

When dataset size and diversity were reduced in Experiment 7, mAP50 decreased to 0.7415 (p-adj. = $2.11 * 10^{-6}$) and mAP50–95 to 0.4448 (p-adj. = 1.17×10^{-5}) (Figure 5). Precision and recall also showed marginal but insignificant reductions to 0.8433 (p-adj. = 0.083) and 0.6835 (p-adj. = 0.083), respectively (Supplementary Figure 9). Augmenting this smaller, limited-diversity dataset from Experiment 7 with synthetic data in Experiment 9 resulted in no significant change in mAP50 (0.7436, p-adj. = 0.95), but mAP50–95 increased significantly (0.4563, p-adj. = 1.1×10^{-4}) (Figure 5). Precision and recall did not change significantly (0.8537, p-adj. = 0.94 and 0.6798, p-adj. = 0.96, respectively) (Supplementary Figure 8).

4 Discussion

Automated detection frameworks for wildlife monitoring using aerial imagery can improve accuracy and reduce manual effort (Gray et al., 2019; Delplanque et al., 2023), accelerating the creation of knowledge needed to study and conserve species at risk.

TABLE 2 Mean average precision at an IoU threshold of 0.50 (mAP50), mean average precision averaged over IoU thresholds from 0.50 to 0.95 (mAP50–95), precision, recall, and F1 score for all 9 experimental detection models evaluated in this study.

Experiment	mAP50	mAP50–95	Precision	Recall	F1 Score
E1[#]	0.7770 ± 0.0068	0.5008 ± 0.0014	**0.8744 ± 0.0276**	0.7005 ± 0.0149	0.7775 ± 0.0071
E2	0.5817 ± 0.0412	0.2778 ± 0.0439	0.6829 ± 0.0806	0.5321 ± 0.0242	0.5974 ± 0.0444
E3	0.7616 ± 0.0092	0.4867 ± 0.0080	0.8579 ± 0.0177	0.6878 ± 0.0121	0.7633 ± 0.0090
E4	0.4070 ± 0.0196	0.1896 ± 0.0114	0.5838 ± 0.0314	0.3779 ± 0.0231	0.4584 ± 0.0222
E5	**0.7782 ± 0.0034**	**0.5014 ± 0.0044**	0.8669 ± 0.0260	**0.7060 ± 0.0186**	**0.7778 ± 0.0073**
E6	0.7464 ± 0.0110	0.4542 ± 0.0059	0.8460 ± 0.0094	0.6888 ± 0.0124	0.7593 ± 0.0075
E7	0.7415 ± 0.0072	0.4448 ± 0.0050	0.8433 ± 0.0214	0.6835 ± 0.0144	0.7547 ± 0.0075
E8	0.7557 ± 0.0080	0.4698 ± 0.0048	0.8543 ± 0.0161	0.6909 ± 0.0189	0.7637 ± 0.0062
E9	0.7436 ± 0.0091	0.4563 ± 0.0056	0.8537 ± 0.0273	0.6798 ± 0.0107	0.7566 ± 0.0107

Values are reported as the mean ± standard deviation across repeated training runs. Bold denotes best values at each metric. Experiment 1 (E1[#]) represents the baseline model used for comparisons.

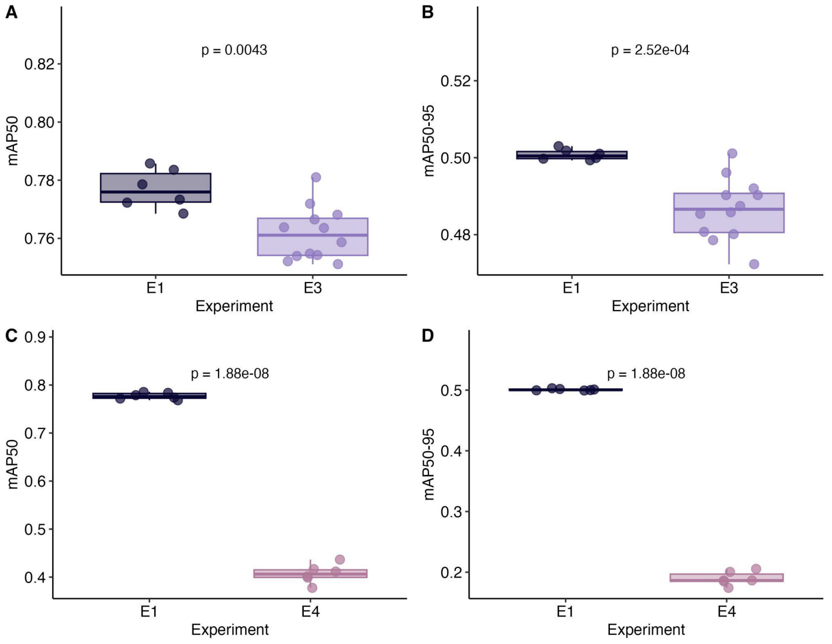

FIGURE 4
Effects of replacing real training images with diffusion-generated synthetic images on detection performance. Mean mAP50 **(A, C)** and mAP50−95 **(B, D)** are shown for models trained using a dataset in which half of the real images were replaced with synthetic images while total training size was held constant (**A, B**; half real + half synthetic vs. real-only [E3 vs. E1]) and models trained using only synthetic images (synthetic-only vs. real-only [**C, D**; E4 vs. E1]). P-values are derived from t-tests and adjusted for multiple tests within comparisons using Holm's correction.

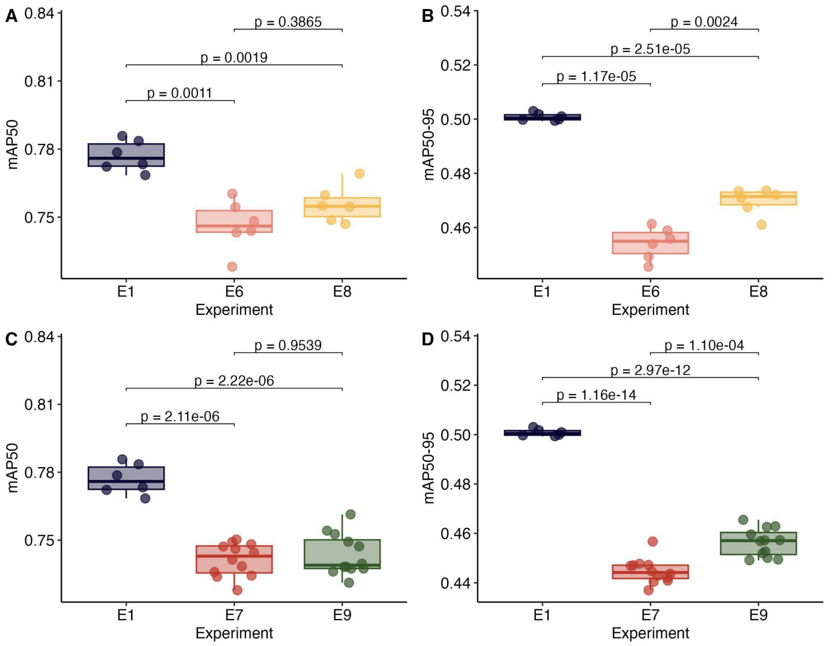

FIGURE 5
Effects of adding diffusion-generated synthetic images to training datasets with reduced real-data diversity and size. Mean mAP50 **(A, C)** and mAP50−95 **(B, D)** are shown for models trained using the full real dataset (real-only baseline [E1]), a reduced-diversity real dataset (fewer flights [E6]), and the same reduced-diversity dataset supplemented with synthetic images (fewer flights + synthetic images [E8]) **(A, B)**, and for models trained using the full real dataset (real-only baseline [E1]), a smaller reduced-diversity real dataset (fewer flights + fewer images [E7]), and the same restricted dataset supplemented with a matched synthetic subset (fewer flights + fewer images + synthetic images [E9]) **(C, D)**. Brackets indicate pairwise comparisons, with p-values derived from t-tests and adjusted for multiple tests within comparisons using Holm's correction.

However, idealized scenarios for model training with large, representative, and diverse training datasets are often unmet. Although synthetic image augmentation has shown promise in improving model performance in other domains (He et al., 2023), its application within wildlife ecology remains underexplored. Here, we evaluated the feasibility of using diffusion-based synthetic imagery generated via LoRA fine-tuning to augment training datasets for humpback whale detection.

4.1 Differences in dataset diversity and influence on detection model accuracy

We observed significantly decreased performance relative to the baseline [E1] when built-in augmentation approaches in the YOLO model were disabled [E2]. We also saw decreases in performance when real images selected from a more limited number of drone flights were used for detection model training [E6]. Both of these experiments had the same number of training instances as the baseline, highlighting that maximizing between-sample diversity alongside leveraging existing augmentation frameworks is essential for optimizing model performance (Wang and Yao, 2009).

The model trained exclusively on synthetic images [E4] performed poorly, and synthetic images in this study exhibited reduced diversity and differed significantly in color distribution compared to real imagery. As a result, models trained solely on synthetic data likely learned features that are not representative of real-world conditions. Doubling the number of instances in the training dataset using synthetic imagery [E5] yielded no significant gains, indicating that synthetic augmentation offered little additional value when real data were already diverse and abundant. These findings reinforce that synthetic imagery should be used to complement rather than replace real data. The optimal balance between real and synthetic training data is likely to be task-specific, and future studies evaluating a broader range of synthetic data proportions are needed to determine how increasing synthetic representation influences detection performance under different data availability conditions (Rankin et al., 2020; Gaur et al., 2023).

4.2 Potential of synthetic imagery to complement real data

Synthetic augmentation showed promise under conditions of data constraint, although substituting real data for synthetic imagery incurred measurable costs. When 50% of real data was replaced with synthetic imagery [E3], mAP50 and mAP50–95 decreased significantly, indicating reduced localization ability relative to the real-only baseline. However, precision and recall did not change significantly, indicating that partial substitution did not measurably alter the model's ability to correctly detect whales or the frequency of false positive detections. This result suggests that limited substitution of real data with synthetic images may preserve detection-level performance when precise localization is less important, but this comes at the cost of reduced average precision performance.

Nonetheless, retaining comparable precision and recall through synthetic augmentation may be desirable in situations where collecting representative real imagery is impractical because of logistical, financial, or environmental constraints, including polar zones, conflict areas, remote islands, or otherwise inaccessible or hazardous regions (Tremblay et al., 2018; Lynch, 2023; Jiguet et al., 2025). Wildlife detection models are also vulnerable to distribution shifts caused by environmental change, including habitat degradation, climate variation, and species range shifts (Taori et al., 2020; Converse et al., 2024). Collecting real training data for all relevant environmental conditions is impractical, particularly for transient states which may not occur often. Consequently, models trained on limited data may fail to generalize to underrepresented environments (Song et al., 2025). Synthetic imagery can help address this limitation by approximating these conditions, enabling training across a broader range of scenarios and supporting model robustness under future environmental change (Pan et al., 2025). In addition, synthetic data may help democratize wildlife detection using high-resolution satellite imagery, where data acquisition is often constrained by high costs and restricted access (Wu et al., 2023). By generating realistic proxy datasets that mimic satellite imagery, diffusion-based approaches can expand data availability and support broader application of satellite detection models (Duporge et al., 2026).

4.3 Enhanced performance under dataset deficiency scenarios

When real training data were limited in size or diversity [E8–E9], synthetic augmentation significantly improved mAP50–95, whereas mAP50, precision, and recall did not differ significantly. This pattern suggests that synthetic image augmentation under data-deficient conditions can improve localization of detected whales in certain cases. Importantly, the mAP50–95 improvement occurred in both data deficiency scenarios despite the limited pixel diversity of the synthetic imagery and the use of LoRA fine-tuning data drawn from a small dataset wholly distinct from the detection model training set of real images. These results demonstrate that LoRA fine-tuning can generate synthetic imagery that provides a measurable downstream benefit under some data-deficient conditions, even if the fine-tuning dataset is independent of the detection model training data, akin to existing transfer learning frameworks (Kumari et al., 2023). However, whether similar benefits occur across other performance metrics, datasets, species, or detection tasks remains to be established.

These findings have potential relevance for boosting detection of species that are hard to detect in aerial images, such as rare or difficult-to-observe species, which often lack sufficient data for effective model development (Beery et al., 2018, Langenkämper et al., 2019). Data deficiency can increase the risk of overfitting, where the model fails to learn distinguishing features of the rare species and instead learns noise present within training images (Janga et al., 2023). Furthermore, wildlife annotations derived from aerial images can often hold a high degree of uncertainty, such as when visibility conditions are poor or distinguishing features are ambiguous (Delisle et al., 2022). These low-certainty annotations are frequently excluded from ML model development efforts (Green et al., 2023), further reducing the available pool of data and potentially biasing models toward more easily detectable or visually distinguishable individuals. Synthetic imagery can address these issues by generating additional training images for rare species and mimicking uncertain visual conditions, such as partially submerged whales, while maintaining a known ground-truth label.

Data deficiency also commonly manifests in the form of class imbalances. Aerial imagery often captures multiple species or behaviors, necessitating robust multi-class detection models (Patton et al., 2023). However, in real-world data, dominant species, conditions, or behaviors are often overrepresented while others are scarce, reducing detection model performance (Leevy et al., 2018). Although we did not explicitly evaluate class imbalance here, the same principle underlying our experiments (augmenting underrepresented training data) could also be applied to minority classes by simulating infrequent behaviors or increasing representation of low-observation species (Jayagopal et al., 2024; Niu et al., 2025). For example, augmenting a camera trap dataset with synthetic images of deer reduced identification errors for this underrepresented class (Beery et al., 2020), illustrating the potential of targeted augmentation to improve model performance.

4.4 Methodological considerations and tradeoffs

Other strategies for synthetic augmentation exist, such as using text-to-image prompts from unoptimized generative diffusion models (Durand et al., 2026), *de novo* training of generative diffusion models (Duporge et al., 2026), or 3D simulation within virtual environments (Gaur et al., 2023). The LoRA fine-tuning approach used in this study offers several practical advantages. First, compared to generic prompt-based generation, fine-tuning enables more consistent control over key image attributes, such as viewing angle or background conditions, without requiring extensive prompt engineering or manual curation (see Supplementary Figure 9 for an example). Durand et al., 2026 reported small increases in performance from models trained on a combination of real and synthetic aerial images of muskox, with synthetic images in their study generated via text-based prompts from OpenAI's image generation diffusion model DALL-E 2. However, this approach yielded only 160 usable images out of 1,000 generated, necessitating extensive prompt engineering to create text prompts suitable for image generation and substantial human filtering. In contrast, the fine-tuning method used in this study requires no manual filtering of generated images and operates using minimal hyperparameter adjustments within the open-source Diffusers framework. Furthermore, this approach is computationally inexpensive. It required only several hours combined for both model fine-tuning and generation of thousands of images using a single NVIDIA A100 GPU, which is readily accessible to any researcher through cloud computing services for only $1 to $5 USD per GPU hour.

Relative to 3D simulation approaches, which often require substantial effort to construct realistic environments and refine object representations (Gaur et al., 2023; Lin et al., 2023), diffusion-based fine-tuning provides a more accessible and efficient alternative using open-source models such as Stable Diffusion XL 1.0. In this approach, the model can directly learn characteristics of the environment and camera angle during fine-tuning, without requiring further parameterization. LoRA fine-tuning also considerably decreases the computational burden associated with training diffusion models *de novo* by leveraging base knowledge from foundation model training to excel in domain-specific tasks (Hu et al., 2021). Other studies evaluating diffusion models for image augmentation have generated millions of images, revealing the scalability of this approach (Azizi et al., 2023).

However, fine-tuning diffusion models depends on the availability of high-quality reference imagery. In cases where these training images are unavailable, prioritizing the collection and annotation of real data remains critical. In addition, we observed homogenization of visual data in the synthetic imagery, with color distributions differing significantly and having reduced diversity compared to real images used for fine-tuning, suggesting LoRA fine-tuning can reduce dataset diversity and constrain model generalization. Finally, diffusion model development demands significant energy and water usage (Li et al., 2025). While using pre-trained diffusion models produces significantly lower emissions than training a base diffusion model, fine-tuning and inference can still be costly. To date, no studies in wildlife ecology have explicitly evaluated how various augmentation strategies differ in computational cost, time investment, and detection performance.

In the present study, all downstream experiments were conducted using a single object detection architecture (YOLOv8s), with humpback whale detection in the Western Antarctic Peninsula serving as a benchmark case study. Different object detection architectures, such as one-stage, two-stage, and transformer-based approaches, can differ in their robustness to distribution shifts and synthetic training data (Zhao et al., 2019; Carion et al., 2020). Consequently, while the workflow presented here for generating and incorporating synthetic imagery is applicable regardless of detection architecture choice, the magnitude of the observed performance gains may be architecture-specific. Likewise, our study did not evaluate transferability across contrasting environmental settings or ecological systems. Environmental conditions such as background characteristics, habitat complexity, and imaging conditions may influence both synthetic image generation and downstream detector performance (Schneider et al., 2020; Corcoran et al., 2021). Determining whether the benefits of synthetic augmentation observed here extend across a broader range of species, ecosystems, environmental conditions, generative models, and object detection architectures is therefore paramount for future applications.

4.5 Avenues for further research

We highlight three areas where further work is needed to advance the application of synthetic imagery in wildlife detection. First, improvements in image generation methods are needed to address the generation of hallucinated features and reduced diversity in synthetic datasets. Although we did not explicitly quantify hallucinations in this study, their presence has been documented previously in diffusion model-generated imagery and may influence downstream model performance (Chen et al., 2025). Distinguishing diffusion-generated artifacts from biologically plausible scenarios in aerial wildlife imagery, such as overlapping or partially occluded animals, is itself challenging and requires standard annotation protocols. Our objective in this study was to evaluate a practical end-to-end workflow requiring minimal manual intervention following image generation, rather than extensive manual curation of synthetic images. Future work should therefore seek to quantify the prevalence of hallucinated features, evaluate how alternative annotation or filtering strategies influence downstream detection model performance, and develop methods that enable more granular control over generated outputs (Sundaram et al., 2025). For example, training multiple fine-tuned generative models, with each one designed to create images reflecting specific environmental conditions, species, or behaviors, may better capture ecological variability and improve generalizability. This could help overcome the homogenization effects from fine-tuning observed in this study and expand the diversity of synthetic training data.

Second, there is a need for improved metrics to evaluate the utility of synthetic imagery. Traditional assessments of image quality often assess visual realism or correlation with real imagery (Heusel et al., 2017; Zhang et al., 2018; Jayasumana et al., 2024) and our initial investigation here into differences among pixel color variation across synthetic and real imagery reveal important differences not immediately obvious to the viewer. For wildlife detection applications, the utility of synthetic images should be defined by the extent to which synthetic data improves downstream model performance, as visually realistic synthetic images do not necessarily provide greater utility for downstream model training (Geng et al., 2024). For wildlife detection applications, synthetic image quality should therefore ultimately be evaluated in relation to downstream task performance. Establishing these relationships will require experiments that systematically quantify or manipulate characteristics of synthetic datasets, including realism, diversity, and presence of hallucinated features, and evaluate their effects on independently trained detection models. Developing metrics which incorporate these characteristics to reliably predict downstream utility for detection represents an important methodological challenge for future research.

Finally, broader validation is required to determine the generalizability of the patterns observed in this benchmark study. The detection performance effects reported here were derived from a single species, ecosystem, generative model, and object detection architecture and therefore cannot be interpreted as universal decision rules for synthetic augmentation. Future studies should test comparable experimental designs across these variables to determine which characteristics govern when synthetic imagery provides benefits or incurs performance costs. Integrating synthetic image generation directly with model training may further improve the efficiency of these workflows. Rather than treating image generation and model development as separate processes, downstream detector performance could guide subsequent rounds of image generation so that synthetic imagery targets specific model weaknesses (Jiang and Wu, 2024).

5 Conclusions

This study provides a robust assessment of how synthetic imagery generated via diffusion model fine-tuning can function as a tool for data augmentation when training wildlife detection models for rare or data-deficient species. Within this case study, replacing 50% of real training images with synthetic data maintained precision and recall but reduced mAP50 and mAP50-95. In contrast, adding synthetic data to real datasets limited in diversity or quantity improved mAP50–95, but mAP50, precision, and recall remained unchanged. Synthetic imagery provided no benefit when diverse real training data were already available and substantially reduced detection performance across all metrics when used without real training imagery. Taken together, these findings demonstrate synthetic imagery generated using diffusion model fine-tuning can be a practical and computationally efficient tool for addressing data limitations in wildlife detection workflows, but its utility depends strongly on the composition of the available training data and the performance objective of interest. Because the quantitative patterns observed here were derived from a single species, ecosystem, generative model, and detection architecture, further validation is needed to determine how these relationships generalize across ecological and computer vision applications. As generative models continue to advance, their integration into wildlife detection workflows will require careful evaluation of both their benefits and limitations.

Data availability statement

A release of the Happy Humpbacks dataset from which real images were sampled is available through the British Antarctic Survey data portal (https://doi.org/10.5285/7a952870-9880-415d-a8ab-194fedf01a26). For this study, all real and synthetic images and annotation files used in detection model training, LoRA model weights, and the fine-tuning imagery and captions are available on Zenodo (https://doi.org/10.5281/zenodo.21401889). All code and execution instructions are available on GitHub (https://github.com/henrysun9074/diffusion-code).

Author contributions

HS: Project administration, Formal analysis, Investigation, Writing – review & editing, Supervision, Validation, Writing – original draft, Data curation, Software, Visualization, Resources, Conceptualization, Methodology. HH: Software, Visualization, Investigation, Data curation, Formal analysis, Validation, Writing – review & editing, Supervision, Conceptualization, Methodology, Writing – original draft, Project administration. JZ: Visualization, Methodology, Investigation, Validation, Software, Writing – review & editing, Formal analysis. MN: Formal analysis, Methodology, Visualization, Software, Investigation, Validation, Writing – review & editing. BB: Conceptualization, Investigation, Writing – review & editing, Resources, Writing – original draft, Supervision, Software, Data curation, Project administration, Validation, Formal analysis, Methodology. DJ: Methodology, Conceptualization, Visualization, Investigation, Project administration, Validation, Funding acquisition, Supervision, Writing – original draft, Data curation, Writing – review & editing, Resources.

Funding

The author(s) declared that financial support was received for this work and/or its publication. HH was supported by the Natural Environment Research Council C-CLEAR doctoral training program (grant no. NE/S007164/1). HS and MN were supported by the Duke University Rachel Carson Scholars Program. HS was also supported by the North Carolina Space Grant Undergraduate Research Scholarship.

Acknowledgments

We thank the Natural Environment Research Council & Duke Marine Robotics and Remote Sensing Lab for enabling collaborations between Duke University and the British Antarctic Survey. This study is a contribution to the Ecosystems Wildlife from Space Project at the British Antarctic Survey as part of the Polar Science for Planet Earth science strategy. Lastly, we thank the two reviewers whose feedback improved the quality of the manuscript.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. Generative AI, namely ChatGPT 5.0, was used to revise code scripts for conciseness and readability, as well as to reformat the README with clearer documentation and instructions for the code repository on Zenodo associated with this manuscript. The authors carefully reviewed and edited all AI-generated outputs and take full responsibility for the final publication content.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2026.1940923/full#supplementary-material

References

Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., and Fleet, D. J. (2023). Synthetic data from diffusion models improves ImageNet classification. *Trans. Mach. Learn. Res.* Available online at: https://openreview.net/forum?id=DlRsoxjyPm (Accessed August 6, 2026).

Beery, S., Liu, Y., Morris, D., Piavis, J., Kapoor, A., Joshi, N., et al. (2020). "Synthetic examples improve generalization for rare classes", in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 863–873. Available online at: https://openaccess.thecvf.com/content_WACV_2020/html/Beery_Synthetic_Examples_Improve_Generalization_for_Rare_Classes_WACV_2020_paper.html (Accessed August 6, 2026).

Beery, S., Van Horn, G., and Perona, P. (2018). "Recognition in terra incognita", in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 456–473. Available online at: https://openaccess.thecvf.com/content_ECCV_2018/html/Beery_Recognition_in_Terra_ECCV_2018_paper.html (Accessed August 6, 2026).

Bierlich, K. C., Hewitt, J., Schick, R. S., Pallin, L., Dale, J., Friedlaender, A. S., et al. (2022). Seasonal gain in body condition of foraging humpback whales along the Western Antarctic Peninsula. *Front. Mar. Sci.* 9. doi: 10.3389/fmars.2022.1036860

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers", in: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, 213–229. doi: 10.1007/978-3-030-58452-8_13

Chen, L., Li, G., Zhang, S., Mao, W., and Zhang, M. (2024). YOLO-SAG: An improved wildlife object detection algorithm based on YOLOv8n. *Ecol. Inf.* 83, 102791. doi: 10.1016/j.ecoinf.2024.102791

Chen, Z., Sun, W., Wu, H., Zhang, Z., Jia, J., Huang, R., et al. (2025). Study of subjective and objective naturalness assessment of AI-generated images. *IEEE Trans. Circuits Syst. Video Technol.* 35, 3573–3588. doi: 10.1109/TCSVT.2024.3509032

Converse, R. L., Lippitt, C. D., Koneff, M. D., White, T. P., Weinstein, B. G., Gibbons, R., et al. (2024). Remote sensing and machine learning to improve aerial wildlife population surveys. *Front. Conserv. Sci.* 5. doi: 10.3389/fcosc.2024.1416706

Corcoran, E., Anschau, M., Sudholz, A., and Hamilton, G. (2021). Automated detection of wildlife using drones: Synthesis, opportunities and constraints. *Methods Ecol. Evol.* 12, 1103–1114. doi: 10.1111/2041-210X.13581

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 10850–10869. doi: 10.1109/TPAMI.2023.3261988

Cuenca, P., and Paul, S. (2023). "Using LoRA for efficient stable diffusion fine-tuning," in *Hugging Face*. Available online at: https://huggingface.co/blog/lora.

Dave, B., Mori, M., Bathani, A., and Goel, P. (2023). "Wild animal detection using YOLOv8", in: *Procedia Computer Science, 3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023)*, 230, 100–111. doi: 10.1016/j.procs.2023.12.065

Deep, B. V., and Dash, R. (2019). "Underwater fish species recognition using deep learning techniques", in: *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*, 665–669. doi: 10.1109/SPIN.2019.8711657

Delisle, Z. J., McGovern, P. G., Dillman, B. G., and Swihart, R. K. (2022). Imperfect detection and wildlife density estimation using aerial surveys with infrared and visible sensors. *Remote Sens. Ecol. Conserv.* 9, 222–234. doi: 10.1002/rse2.305

Delplanque, A., Lamprey, R., Foucher, S., Théau, J., and Lejeune, P. (2023). Surveying wildlife and livestock in Uganda with aerial cameras: Deep learning reduces the workload of human interpretation by over 70%. *Front. Ecol. Evol.* 11. doi: 10.3389/fevo.2023.1270857

Dhariwal, P., and Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Adv. Neural Inf. Process. Syst.* 34, 8780–8794. Available online at: https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html (Accessed August 6, 2026).

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., et al. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mach. Intell.* 5, 220–235. doi: 10.1038/s42256-023-00626-4

Duporge, I., Wu, Z., Black, M. E., Wolf, S., Albi, A., Watt, C. A., et al. (2026). Enhancing satellite-based detection of marine fauna using synthetic data. *Ecol. Inf.* 95, 103767. doi: 10.1016/j.ecoinf.2026.103767

Durand, S., Foucher, S., Delplanque, A., Taillon, J., and Théau, J. (2026). Lacking data? No worries! How synthetic images can alleviate image scarcity in wildlife surveys: A case study with muskox (Ovibos moschatus). *Remote Sens. Ecol. Conserv.* n/a. doi: 10.1002/rse2.70056

El-Kaddoury, M., Mahmoudi, A., and Himmi, M. M. (2019). "Deep generative models for image generation: A practical comparison between variational autoencoders and generative adversarial networks," in *Mobile, Secure, and Programmable Networking*. Eds. É. Renault, S. Boumerdassi, C. Leghris and S. Bouzefrane (Cham: Springer International Publishing), 1–8. doi: 10.1007/978-3-030-22885-9_1

Gaur, A., Liu, C., Lin, X., Karapetyan, N., and Aloimonos, Y. (2023). "Whale detection enhancement through synthetic satellite images," in *Oceans 2023 - Mts/Ieee U.s. Gulf Coast*, 1–7. doi: 10.23919/OCEANS52994.2023.10337400

Geng, S., Hsieh, C.-Y., Ramanujan, V., Wallingford, M., Li, C.-L., Koh, P. W., et al. (2024). The unmet promise of synthetic training images: Using retrieved real images performs better. *Adv. Neural Inf. Process. Syst.* 37, 7902–7929. doi: 10.52202/079017-0254

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 27. Available online at: https://papers.nips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html (Accessed August 6, 2026).

Gray, P. C., Bierlich, K. C., Mantell, S. A., Friedlaender, A. S., Goldbogen, J. A., and Johnston, D. W. (2019). Drones and convolutional neural networks facilitate automated and accurate cetacean species identification and photogrammetry. *Methods Ecol. Evol.* 10, 1490–1500. doi: 10.1111/2041-210X.13246

Gray, P. C., Fleishman, A. B., Klein, D. J., McKown, M. W., Bézy, V. S., Lohmann, K. J., et al. (2018). A convolutional neural network for detecting sea turtles in drone imagery. *Methods Ecol. Evol.* 10, 345–355. doi: 10.1111/2041-210X.13132

Green, K. M., Virdee, M. K., Cubaynes, H. C., Aviles-Rivero, A. I., Fretwell, P. T., Gray, P. C., et al. (2023). Gray whale detection in satellite imagery using deep learning. *Remote Sens. Ecol. Conserv.* 9, 829–840. doi: 10.1002/rse2.352

Guirado, E., Tabik, S., Rivas, M. L., Alcaraz-Segura, D., and Herrera, F. (2019). Whale counting in satellite and aerial images with deep learning. *Sci. Rep.* 9, 14259. doi: 10.1038/s41598-019-50795-9

He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., et al. (2023). "IS SYNTHETIC DATA FROM GENERATIVE MODELS READY FOR IMAGE RECOGNITION?", in: *The Eleventh International Conference on Learning Representations*. Available online at: https://openreview.net/forum?id=nUmCcZ5RKF (Accessed August 6, 2026).

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Adv. Neural Inf. Process. Syst.* 30. Available online at: https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html (Accessed August 6, 2026).

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* 33, 6840–6851. Available online at: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html (Accessed August 6, 2026).

Houliston, H. R., Cheng, Y., Johnston, D. W., Larsen, G. D., Friedlaender, A. S., Fretwell, P. T., et al. (2026). *Happy Humpbacks: A machine learning-ready drone dataset for whale detection model development (Version 1.0, p. 10566 files, 55.8 GB)* (NERC EDS UK Polar Data Centre). doi: 10.5285/7A952870-9880-415D-A8AB-194FEDF01A26

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al. (2021). "LoRA: low-rank adaptation of large language models," in *The Tenth International Conference on Learning Representations*. Available online at: https://openreview.net/forum?id=nZeVKeeFYf9 (Accessed August 6, 2026).

Jabbar, A., Li, X., and Omar, B. (2022). A survey on generative adversarial networks: Variants, applications, and training. *ACM Comput. Surv.* 54, 1–49. doi: 10.1145/3463475

Janga, B., Asamani, G. P., Sun, Z., and Cristea, N. (2023). A review of practical AI for remote sensing in Earth sciences. *Remote Sens.* 15, Article 16. doi: 10.3390/rs15164112

Jayagopal, P., Purushothaman Janaki, K., Mohan, P., Kondapaneni, U. B., Periyasamy, J., Mathivanan, S. K., et al. (2024). A modified generative adversarial networks with Yolov5 for automated forest health diagnosis from aerial imagery and Tabu search algorithm. *Sci. Rep.* 14, 4814. doi: 10.1038/s41598-024-54399-w

Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., and Kumar, S. (2024). "Rethinking FID: Towards a better evaluation metric for image generation", in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9307–9315. Available online at: https://openaccess.thecvf.com/content/CVPR2024/html/Jayasumana_Rethinking_FID_Towards_a_Better_Evaluation_Metric_for_Image_Generation_CVPR_2024_paper.html (Accessed August 6, 2026).

Jiang, L., and Wu, L. (2024). Enhanced Yolov8 network with Extended Kalman Filter for wildlife detection and tracking in complex environments. *Ecol. Inf.* 84, 102856. doi: 10.1016/j.ecoinf.2024.102856

Jiguet, F., Mayrose, A., Piha, M., Efrat, R., Seimola, T., Sapir, N., et al. (2025). GNSS spoofing in conflict zones disrupts wildlife tracking and hampers research and conservation efforts. *Nat. Commun.* 16, 1199. doi: 10.1038/s41467-025-56630-2

Johnston, D. W. (2019). Unoccupied aircraft systems in marine science and conservation. *Annu. Rev. Mar. Sci.* 11, 439–463. doi: 10.1146/annurev-marine-010318-095323

Kingma, D. P., and Welling, M. (2022). Auto-encoding variational bayes. *arXiv*. doi: 10.48550/arXiv.1312.6114

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and Zhu, J.-Y. (2023). "Multi-concept customization of text-to-image diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1931–1941. Available online at: https://openaccess.thecvf.com/content/CVPR2023/html/Kumari_Multi-Concept_Customization_of_Text-to-Image_Diffusion_CVPR_2023_paper.html (Accessed August 6, 2026).

Langenkämper, D., van Kevelaer, R., and Nattkemper, T. W. (2019). "Strategies for tackling the class imbalance problem in marine image classification," in *Pattern Recognition and Information Forensics*. Eds. Z. Zhang, D. Suter, Y. Tian, A. Branzan Albu, N. Sidère and H. Jair Escalante (Cham: Springer International Publishing), 26–36. doi: 10.1007/978-3-030-05792-3_3

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., and Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *J. Big Data* 5, 42. doi: 10.1186/s40537-018-0151-6

Li, P., Yang, J., Islam, M. A., and Ren, S. (2025). Making AI less "thirsty. *Commun. ACM* 68, 54–61. doi: 10.1145/3724499

Lin, X., Sanket, N. J., Karapetyan, N., and Aloimonos, Y. (2023). "OysterNet: Enhanced oyster detection using simulation", in: *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 5170–5176. doi: 10.1109/ICRA48891.2023.10160830

Liu, Y., Yue, J., Xia, S., Ghamisi, P., Xie, W., and Fang, L. (2024). Diffusion models meet remote sensing: Principles, methods, and perspectives. *IEEE Trans. Geosci. Remote Sens.* 62, 1–22. doi: 10.1109/TGRS.2024.3464685

Lu, C.-Y., Arcega Rustia, D. J., and Lin, T.-T. (2019). "Generative adversarial network based image augmentation for insect pest classification enhancement", in: *IFAC-PapersOnLine, 6th IFAC Conference on Sensing, Control and Automation Technologies For. Agriculture AGRICONTROL 2019*, 52, 1–5. doi: 10.1016/j.ifacol.2019.12.406

Lynch, H. J. (2023). Satellite remote sensing for wildlife research in the polar regions. *Mar. Technol. Soc J.* 57, 43–50. doi: 10.4031/MTSJ.57.3.1

Michener, W. K., and Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. *Trends Ecol. Evol.* 27, 85–93. doi: 10.1016/j.tree.2011.11.016

Muhammad, A., Salman, Z., Lee, K., and Han, D. (2023). Harnessing the power of diffusion models for plant disease image augmentation. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1280496

Müller-Franzes, G., Niehues, J. M., Khader, F., Arasteh, S. T., Haarburger, C., Kuhl, C., et al. (2023). A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Sci. Rep.* 13, 12098. doi: 10.1038/s41598-023-39278-0

Mumuni, A., and Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array* 16, 100258. doi: 10.1016/j.array.2022.100258

Niu, Z., Li, Y., Gong, Y., Zhang, B., He, Y., Zhang, J., et al. (2025). Multi-class guided GAN for remote-sensing image synthesis based on semantic labels. *Remote Sens.* 17, 2. doi: 10.3390/rs17020344

Pan, J., Lei, S., Fu, Y., Li, J., Liu, Y., Sun, Y., et al. (2025). EarthSynth: Generating informative Earth observation with diffusion models. *arXiv*. doi: 10.48550/arXiv.2505.12108

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32. . Available online at: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Patton, P. T., Cheeseman, T., Abe, K., Yamaguchi, T., Reade, W., Southerland, K., et al. (2023). A deep learning approach to photo–identification demonstrates high performance on two dozen cetacean species. *Methods Ecol. Evol.* 14, 2611–2625. doi: 10.1111/2041-210X.14167

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., et al. (2023). "SDXL: Improving latent diffusion models for high-resolution image synthesis", in: *The Twelfth International Conference on Learning Representations*. Available online at: https://openreview.net/forum?id=di52zR8xgf.

Rafiq, K., Beery, S., Palmer, M. S., Harchaoui, Z., and Abrahms, B. (2025). Generative AI as a tool to accelerate the field of ecology. *Nat. Ecol. Evol.* 9, 378–385. doi: 10.1038/s41559-024-02623-1

Rankin, D., Black, M., Bond, R., Wallace, J., Mulvenna, M., and Epelde, G. (2020). Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR Med. Inf.* 8, e18910. doi: 10.2196/18910

R Core Team (2025). R: The R Project for Statistical Computing. Available online at: https://www.r-project.org/.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016).You only look once: Unified, real-time object detection. Available online at: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html (Accessed August 6, 2026).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). "High-resolution image synthesis with latent diffusion models", in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Roy, A. M., Bhaduri, J., Kumar, T., and Raj, K. (2023). WilDect-YOLO: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. *Ecol. Inf.* 75, 101919. doi: 10.1016/j.ecoinf.2022.101919

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2023). "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22500–22510. Available online at: https://openaccess.thecvf.com/content/CVPR2023/html/Ruiz_DreamBooth_Fine_Tuning_Text-to-Image_Diffusion_Models_for_Subject-Driven_Generation_CVPR_2023_paper.html (Accessed August 6, 2026).

Schneider, S., Greenberg, S., Taylor, G. W., and Kremer, S. C. (2020). Three critical factors affecting automated image species recognition performance for camera traps. *Ecol. Evol.* 10, 3503–3517. doi: 10.1002/ece3.6147

Seymour, A. C., Dale, J., Hammill, M., Halpin, P. N., and Johnston, D. W. (2017). Automated detection and enumeration of marine wildlife using unmanned aircraft systems (UAS) and thermal imagery. *Sci. Rep.* 7, 45127. doi: 10.1038/srep45127

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 1–48. doi: 10.1186/s40537-019-0197-0

Song, S., Tang, Y., and Qin, R. (2025). Synthetic data matters: Retraining with geo-typical synthetic labels for building detection. *IEEE Trans. Geosci. Remote Sens.* 63, 1–13. doi: 10.1109/TGRS.2025.3593864

Sundaram, S., Chae, J., Tian, Y., Beery, S., and Isola, P. (2025). "Personalized representation from personalized generation", in: *International Conference on Learning Representations*, 2025, 92572–92606.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. (2020). "Measuring robustness to natural distribution shifts in image classification," in *Advances in Neural Information Processing Systems*, vol. 33. , 18583–18599. Available online at: https://proceedings.neurips.cc/paper/2020/hash/d8330f857a17c53d217014ee776bfd50-Abstract.html (Accessed August 6, 2026).

Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., et al. (2018). "Training deep networks with synthetic data: Bridging the reality gap by domain randomization", in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 969–977. Available online at: https://openaccess.thecvf.com/content_cvpr_2018_workshops/w14/html/Tremblay_Training_Deep_Networks_CVPR_2018_paper.html (Accessed August 6, 2026).

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., et al. (2022). Diffusers: State-of-the-Art Diffusion Models. Available online at: https://github.com/huggingface/diffusers.

Wada, K. (2026). *Labelme: Image Polygonal Annotation With Python*. doi: 10.5281/zenodo.5711226

Wang, S., and Yao, X. (2009). "Diversity analysis on imbalanced data sets by using ensemble models", in: *2009 IEEE Symposium on Computational Intelligence and Data Mining*, 324–331. doi: 10.1109/CIDM.2009.4938667

Weinstein, B. G. (2018). A computer vision for animal ecology. *J. Anim. Ecol.* 87, 533–545. doi: 10.1111/1365-2656.12780

Wu, Z., Zhang, C., Gu, X., Duporge, I., Hughey, L. F., Stabach, J. A., et al. (2023). Deep learning enables satellite-based monitoring of large populations of terrestrial mammals across heterogeneous landscape. *Nat. Commun.* 14, 3072. doi: 10.1038/s41467-023-38901-y

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). "The unreasonable effectiveness of deep features as a perceptual metric", in: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595. doi: 10.1109/CVPR.2018.00068

Zhao, Z.-Q., Zheng, P., Xu, S.-T., and Wu, X. (2019). Object detection with deep learning: A review. *IEEE Trans. Neural Networks Learn. Syst.* 30, 3212–3232. doi: 10.1109/TNNLS.2018.2876865