

Closing the Rare-Species Detection Gap in UAE Camera Traps with Diffusion-Generated Data

Abdu Sallouh¹^[0009-0008-9785-9853], Aariz Faridi¹, Khumam Alzeem¹, Mohamad Dabbagh¹, Ali Reza Sajun¹^[0000-0002-1048-5633], Imran Zualkernan¹^[0000-0003-1270-3005], Altaf Habib², and Andrew Gardner²

¹ ¹ American University of Sharjah, UAE
{b00087818, b00086502, b00085319, b00101218, b00068908, izualkernan}@aus.edu
² ² Emirates Nature-WWF, UAE
{ahabib, agardner}@enwwf.ae

Abstract. In biodiversity monitoring, the species most critical to conservation outcomes are often those least represented in training data. This study examines whether diffusion-based synthesis can improve detection of such tail classes in long-tailed UAE camera-trap data without additional field collection. DreamBooth Stable Diffusion v1-5 was fine-tuned to generate approximately 700 images per under-represented class (Camel, Cat, *Ammomanes deserti*, *Emberiza striolata*, *Oenanthe albonigra*) and combined with real data under uniform and non-uniform sampling at 1:1, 1:2, and 1:4 ratios with use of YOLOv6 Lite-S. Results were class-dependent: Cat F1-score increased by up to 0.36, Camel reached F1-score of 0.86 at 1:4 non-uniform ratio, whereas the smallest bird class (*Ammomanes deserti*) consistently deteriorated, dropping to F1 scores as low as 0.04, while other bird species showed mixed results. High FID scores ranging from 634 to 1362 indicate a realism and domain-match gap that constrains transfer; gains correlated more strongly with object scale, visibility, and background congruence than with synthetic volume alone. These findings suggest diffusion-assisted rebalancing can reduce annotation cost and selectively enhance sensitivity to scarce species that drive conservation decisions, but effectiveness is not uniform across species. A class-aware synthesis protocol, stricter exemplar filtering, habitat- and pose-aware prompt design, is recommended to narrow the realism gap and stabilize benefits for small-bodied targets.

Keywords: Camera trap detection, Long-tailed distribution, Diffusion models, Rare species, Data augmentation, Object detection

1 Introduction

Biodiversity monitoring increasingly relies on camera traps to estimate species activity, but these datasets are typically long-tailed, with common species over-represented and rare species sparsely sampled, a pattern quantified in recent

camera-trap studies and benchmarks (Cunha et al., 2023; He et al., 2023). Manual annotation at scale is costly and slow, and while modern object detectors have improved throughput, rare-class performance remains limited. In the United Arab Emirates (desert and mangrove systems, strong day/night variation), these challenges are amplified. Improving rare-class detection without additional field-work would reduce annotation burden and increase the value of camera-trap programs for conservation decisions. Our objective is to test whether class-conditional diffusion-generated images improve rare-class detection in camera-trap data without additional field collection, and to characterize when synthesis helps or harms. Specifically, we ask whether adding diffusion-generated samples (Trabucco et al., 2024; Chen et al., 2022) raises rare-class F1-score over a real-only baseline; how mixing ratios (1:1, 1:2, 1:4) and sampling policies (uniform vs. non-uniform) influence rare-class F1-score and overall mAP; to what extent object scale/visibility and background context alignment mediate outcomes; and how synthetic–real similarity relates to gains or regressions. We consider the approach successful if rare-class improvements are achieved without material loss in head-class performance or overall mAP.

We evaluate these questions on an Emirates Nature–WWF camera-trap dataset from the Hajar Mountains. The distribution is strongly long-tailed, and imagery reflects typical field constraints: day/IR-night bursts, motion triggers, small or partially occluded targets, and frequent empty frames. We generate synthetic images for five tail classes, namely: Camel, Cat, *Ammomanes deserti* (AD), *Emberiza striolata* (ES), and *Oenanthe albonigra* (OA), using DreamBooth Stable Diffusion (v1-5), then train YOLOv6 Lite-S on mixtures of real and synthetic data under uniform and non-uniform sampling at 1:1, 1:2, and 1:4 ratios. Evaluation emphasizes rare-class F1-score alongside overall mAP, with FID used to quantify synthetic–real similarity.

2 Related Work

Approaches to long-tailed detection act on two levers: loss adjustments that give rare classes more influence, such as focal loss, class-balanced loss, and logit adjustment (Lin et al., 2017b; Cui et al., 2019; Cao et al., 2019; Menon et al., 2021); and data/sampler designs that rebalance exposure through oversampling, selective down-sampling, or decoupled classifier heads (Cheng et al., 2020; Tan et al., 2020; Li et al., 2020).

Small targets present additional challenges due to weak features and disrupted assignment under occlusion, clutter, and IR noise. Multi-scale backbones such as Feature Pyramid Networks, super-resolution preprocessing, and domain adaptation for low-light settings address some of these issues (Lin et al., 2017a; Haris et al., 2018; Du et al., 2024), though scale and background congruence remain dominant failure modes in camera-trap imagery (Koh et al., 2021).

Diffusion models have improved control over synthetic data generation through text conditioning and personalization. DreamBooth adapts pretrained models from few exemplars while maintaining species features, and extensions such as ControlNet help preserve composition and viewpoint (Ruiz et al., 2023; Zhang

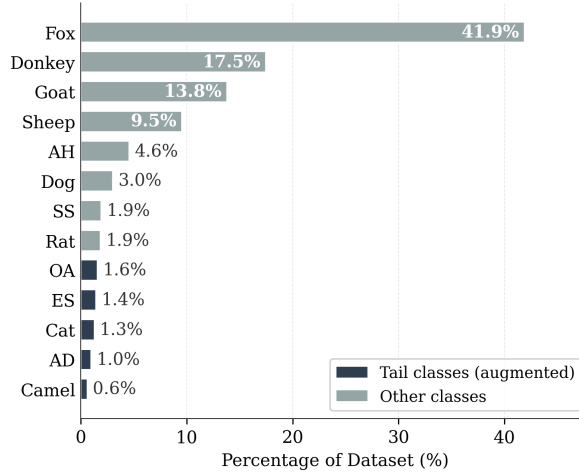


Fig. 1. Class distribution of animal species in the Emirates Nature-WWF dataset.

et al., 2023). For detection, gains concentrate when synthetic samples match habitat and maintain realistic scale, while regressions appear with background mismatch or texture drift. Copy-paste over real backgrounds remains a strong baseline, and filtering through removing near-duplicates, enforcing per-class caps, or checking size/viewpoint consistency is as important as generation itself (Zhao et al., 2023; Fang et al., 2024; Lin et al., 2023).

3 Methodology

3.1 Dataset Description

The dataset employed in this study originates from Emirates Nature-WWF and comprises camera-trap imagery collected across the Hajar Mountains of the United Arab Emirates. It contains approximately 66,851 images annotated into 15 classes, including both day and night captures at a native resolution of 1920×1440 pixels. Each image is accompanied by bounding-box annotations in COCO JSON format specifying class ID, object coordinates, and metadata such as capture time and location. The distribution follows a pronounced long-tailed pattern typical of ecological datasets (Fig. 1), dominated by a “Ghost” (empty-frame) class and several head species (e.g., Fox, Donkey), while rare species such as Camel, Cat, AD, ES, and OA are severely under-represented. These tail categories often contain fewer than 200 samples and present additional challenges such as small body size relative to frame resolution, partial occlusion, camouflage against desert terrain, and frequent low-illumination infrared captures.

3.2 Data Preparation and Tail-Class Selection

Prior to model training, the dataset was cleaned and reformatted to ensure compatibility with both the generative and detection pipelines. The subset used in

this study comprises 8,355 labeled images, following the same 60–20–20 train, validation, and test split established in prior work on the Emirates Nature–WWF dataset. Each image was visually inspected to remove invalid or unusable samples, such as frames containing only partial animal features (e.g., tails or blurred silhouettes) or empty motion-triggered shots. Bounding-box annotations provided in COCO JSON format were parsed to generate YOLO-compatible .txt labels for each class, listing object coordinates normalized to image dimensions. All images were resized automatically to 640×480 pixels during YOLOv6 preprocessing to match the detector’s input requirements, while diffusion-based image generation was conducted on downscaled 512×512 versions to align with Stable Diffusion’s native resolution. For targeted rebalancing, five tail classes were identified for generative augmentation based on their low sample counts and ecological relevance: Camel, Cat, AD, ES, and OA. Each selected class was reviewed manually to retain clear, centered, and well-lit animal instances suitable for fine-tuning the DreamBooth Stable Diffusion model.

3.3 Synthetic Data Generation

DreamBooth fine-tuning was applied to the “runwayml/stable-diffusion-v1-5” checkpoint for each of the five tail classes. Input images were manually screened to retain only clear, full-body animal representations, removing frames with partial visibility or missing subjects. Images were resized to 512×512 pixels to match Stable Diffusion’s native resolution.

Following Patil and Hu (2022), a learning rate of 1×10^{-6} and 2,000 training steps were used; higher learning rates or extended training caused overfitting, visible as unnatural textures and background color shifts. DreamBooth was chosen for its capacity to adapt pretrained diffusion models from limited examples while maintaining species features and environmental realism.

3.4 Rebalancing Strategies

After generating the synthetic images through DreamBooth fine-tuning, a structured rebalancing approach was implemented to correct the imbalance observed among the five rare species, while preserving the natural variability of the dataset. As such, uniform and non-uniform sampling are adopted. In the uniform configuration, all tail classes were equalized to a similar total number of images after augmentation, ensuring that each class contributed equally to the training process. In the non-uniform configuration, the relative proportions between classes were maintained, but each class was supplemented with a number of generated images proportional to its original sample size. Experiments were conducted across three augmentation ratios: 1:1, 1:2, and 1:4, corresponding respectively to adding one, two, and four generated images for every real image in the tail classes. Following data preparation, all rebalanced sets were used to fine-tune the YOLOv6-Lite-S model for 40 epochs with a batch size of 200 and an image resolution of 640×480 on an Nvidia A100 GPU. Before integration, all generated samples were visually screened to remove duplicates, unrealistic backgrounds, and incomplete subjects. The resulting filtered synthetic subsets were then ap-

pended to the original training data under each ratio configuration. The resulting image counts per class and configuration are summarized in Table 1.

Table 1. Distribution of training images per tail class across uniform and non-uniform augmentation strategies at different augmentation ratios.

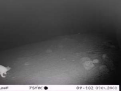
Class	Real	Uniform			Non-Uniform		
		1:1	1:2	1:4	1:1	1:2	1:4
AD	86	212	317	528	172	258	430
Camel	50	212	317	528	100	150	250
Cat	128	212	317	528	256	384	640
ES	109	212	317	528	218	327	545
OA	155	212	317	528	310	465	775

4 Results

4.1 Diffusion Image Quality and Realism

The DreamBooth Stable Diffusion model was fine-tuned on the five tail classes (Camel, Cat, AD, ES, and OA) to generate additional training samples. A qualitative comparison between real and generated samples is shown in Table 2, which presents side-by-side examples for each class. The generated images successfully reproduce the lighting conditions, terrain texture, and infrared style typical of the dataset. However, variation in generation quality was observed between species. Larger and visually distinct animals, such as Camel, produced realistic outputs closely resembling the original captures, while smaller or partially obscured animals, such as AD and ES, exhibited minor distortions and texture inconsistencies.

Table 2. Sample of Generated Images

Image Type	AD	Camel	Cat	ES	OA
Real Image					
Generated Image					

Quantitative evaluation using the Fréchet Inception Distance (FID) further supports these observations. As summarized in Table 3, Camel achieved the lowest FID score, indicating high similarity to real images, while Cat had the highest FID, it still yielded the largest F1 gain, suggesting that FID alone did not predict benefits in detection scores.

Table 3. FID scores for generated images across tail classes (lower is better).

Class	AD	ES	OA	Camel	Cat
FID ($\downarrow$)	1202.07	858.99	743.36	634.09	1361.89

4.2 Baseline Model Performance

Before introducing synthetic data, the YOLOv6-Lite-S model was fine-tuned on the original unbalanced subset. The results, summarized in Table 4, serve as the baseline against which all diffusion-augmented configurations were compared. Tail classes that are later targeted by diffusion augmentation are highlighted in bold.

Table 4. Baseline YOLOv6-Lite-S performance on the original WWF training subset. Tail classes targeted for diffusion augmentation are in bold.

Class	Labels	P@0.5	R@0.5	F1@0.5	mAP@0.5
All	3247	0.677	0.464	0.487	0.505
Rat	65	0.634	0.0615	0.112	0.168
AD	35	0.594	0.0571	0.104	0.098
Ammoperdix heyi	165	0.668	0.476	0.556	0.522
Camel	18	0.743	0.833	0.786	0.811
Cat	44	1.000	0.000	0.000	0.071
Dog	102	0.620	0.725	0.668	0.707
Donkey	595	0.892	0.844	0.867	0.914
ES	51	0.479	0.270	0.345	0.317
Fox	1225	0.875	0.914	0.894	0.939
Goat	444	0.738	0.752	0.745	0.795
OA	56	0.313	0.196	0.241	0.229
Sheep	378	0.842	0.762	0.800	0.833
Spilopelia senegalensis	69	0.399	0.145	0.213	0.154

4.3 Uniform Sampling Configurations

Three uniform augmentation ratios were tested (1:1, 1:2, and 1:4), where every real image from the tail classes was supplemented with diffusion-generated images and per-class totals were equalized. As shown in Tables 5 and 6, the 1:1 configuration produced moderate improvements in several classes, with Cat achieving the largest gain (+0.20 F1, +0.09 mAP) and OA showing minor improvement, while AD experienced a small decrease. At the 1:2 ratio, Camel remained above baseline while Cat continued to improve in mAP@0.5, whereas bird species declined further, with AD dropping to an F1-score of 0.04. At the 1:4

Table 5. Comparison of F1-scores across all augmentation strategies. **Baseline** refers to the real-only training set. The best result per class is bolded.

Class	Base	Uniform			Non-Uniform		
	Real	1:1	1:2	1:4	1:1	1:2	1:4
AD	0.104	0.087	0.044	0.068	0.083	0.040	0.082
Camel	0.786	0.744	0.809	0.669	0.683	0.770	0.860
Cat	0.000	0.203	0.269	0.361	0.188	0.211	0.288
ES	0.345	0.349	0.332	0.467	0.340	0.311	0.423
OA	0.241	0.313	0.278	0.230	0.265	0.345	0.242
All	0.487	0.494	0.498	0.509	0.487	0.499	0.516

Table 6. Comparison of mAP@0.5 across all augmentation strategies.

Class	Base	Uniform			Non-Uniform		
	Real	1:1	1:2	1:4	1:1	1:2	1:4
AD	0.098	0.050	0.039	0.046	0.034	0.039	0.052
Camel	0.811	0.852	0.848	0.827	0.814	0.854	0.858
Cat	0.071	0.156	0.207	0.193	0.091	0.174	0.150
ES	0.317	0.227	0.264	0.371	0.279	0.240	0.332
OA	0.229	0.234	0.229	0.226	0.210	0.257	0.219
All	0.505	0.497	0.503	0.509	0.490	0.500	0.507

ratio, Cat and ES maintained moderate improvements, though Camel’s F1-score dropped and AD continued to underperform.

4.4 Non-Uniform Sampling Configurations

Non-uniform sampling was also tested to evaluate how preserving natural class proportions influences outcomes. In these configurations, each tail class was supplemented proportionally rather than equalized. At the 1:1 ratio, results resembled the 1:1 uniform case, with Cat again showing the largest F1-score gain (+0.19) and OA improving modestly, while AD exhibited persistent degradation and ES remained relatively flat across most configurations. The 1:2 ratio followed similar trends, with Camel stabilizing at an F1-score around 0.77. At the 1:4 ratio, non-uniform augmentation yielded the strongest results for Camel, which reached an F1-score of 0.86 and mAP@0.5 above 0.85, while AD remained the weakest performer across all configurations.

5 Discussion

The experiments indicate that diffusion-based augmentation can improve detection for larger animals such as Camel and Cat, but provides limited benefit for the smallest bird species such as AD, while ES showed moderate improvement at

higher augmentation ratios. Because these birds span only a few pixels, even minor shape or texture distortions can shift synthetic samples off the true distribution. This sensitivity is reflected in the consistently high FID scores (634–1362), which confirm a substantial realism gap constraining transfer. Notably, increasing synthetic volume alone did not guarantee improvement; instead, gains depended more on object scale and background congruence than on augmentation ratio. Furthermore, uniform and non-uniform sampling produced similar overall trends, suggesting that class-specific properties dominate over the choice of balancing strategy.

6 Conclusion and Future Work

This study evaluated whether diffusion-generated images can improve rare-species detection in long-tailed UAE camera-trap data by fine-tuning DreamBooth Stable Diffusion on five under-represented classes and integrating synthetic samples under multiple mixing strategies. The results show that diffusion-based augmentation can improve detection for larger and visually distinctive species, such as Camel and Cat, but offers limited or inconsistent benefit for small birds, with the smallest species (AD) consistently underperforming despite augmentation. Although generative synthesis reduced annotation constraints for some classes, its effectiveness was not uniform. Gains were strongly tied to how well the generated images matched the visual and contextual characteristics of real camera-trap frames. This suggests that successful augmentation in ecological settings will require more selective, context-aware, and class-specific generation procedures rather than uniform scaling of synthetic data. Overall, diffusion-assisted rebalancing shows promise as a complementary tool for camera-trap pipelines, particularly for species with clear structure and limited field samples, and provides a foundation for more advanced and controlled generative approaches in future work.

Future work can explore more controlled forms of diffusion-based generation to reduce the realism gap observed in this study. Recent approaches such as ControlNet (Zhang et al., 2023) introduce spatial guidance during synthesis and may help preserve object boundaries, viewpoints, and habitat cues while still adding variation in pose or scale. The challenges faced by the smaller bird classes suggest that foreground appearance alone is not sufficient. A practical next step is to combine diffusion outputs with hybrid augmentation strategies that keep real backgrounds intact. Conditioning future generations on habitat type, illumination, or camera viewpoint could help create more robust models that perform more consistently across unseen locations or seasons.

Acknowledgement

The authors would like to express their gratitude to Emirates Nature–WWF, UAE, for their valuable assistance and generous provision of camera trap image data.

References

- Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. 2019.
- Dong Chen, Xinda Qi, Yu Zheng, Yuzhen Lu, and Zhaojian Li. Deep Data Augmentation for Weed Recognition Enhancement, 2022.
- Bowen Cheng, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas Huang, and Humphrey Shi. Decoupled Classification Refinement: Hard False Positive Suppression for Object Detection, 2020.
- Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-Balanced Loss Based on Effective Number of Samples. 2019.
- Fagner Cunha, Eulanda M. Dos Santos, and Juan G. Colonna. Bag of tricks for long-tail visual recognition of animal species in camera-trap images. *Ecological Informatics*, 76:102060, 2023. doi: 10.1016/j.ecoinf.2023.102060.
- Zhipeng Du, Miaoqing Shit, and Jiankang Deng. Boosting Object Detection with Zero-Shot Day-Night Domain Adaptation. In *IEEE/CVF CVPR*, pages 12666–12676, 2024. doi: 10.1109/CVPR52733.2024.01204.
- Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. Data Augmentation for Object Detection via Controllable Diffusion Models. In *IEEE/CVF WACV*, pages 1246–1255, 2024. doi: 10.1109/WACV57701.2024.00129.
- Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-Driven Super Resolution: Object Detection in Low-resolution Images, 2018.
- Wentong He, Ze Luo, Xinyu Tong, Xiaoyi Hu, Can Chen, and Zufe Shu. Long-Tailed Metrics and Object Detection in Camera Trap Datasets. *Applied Sciences*, 13:6029, 2023. doi: 10.3390/app13106029.
- Pang Wei Koh, Shiori Sagawa, et al. WILDS: A Benchmark of in-the-Wild Distribution Shifts. 2021.
- Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming Classifier Imbalance for Long-Tail Object Detection With Balanced Group Softmax. In *IEEE/CVF CVPR*, pages 10988–10997, 2020. doi: 10.1109/CVPR42600.2020.01100.
- Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the Power of Synthetic Data on Few-shot Object Detection. In *IEEE/CVF CVPRW*, pages 638–647, 2023. doi: 10.1109/CVPRW59228.2023.00071.
- Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *IEEE CVPR*, pages 936–944, 2017a. doi: 10.1109/CVPR.2017.106.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. *IEEE TPAMI*, 2017b.
- Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021.

- Suraj Patil and Edward Hu. DreamBooth training guide. Hugging Face Blog, 2022.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE/CVF CVPR*, pages 22500–22510, 2023. doi: 10.1109/CVPR52729.2023.02155.
- Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization Loss for Long-Tailed Object Recognition. In *IEEE/CVF CVPR*, pages 11659–11668, 2020. doi: 10.1109/CVPR42600.2020.01168.
- Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective Data Augmentation with Diffusion Models. 2024.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *IEEE/CVF ICCV*, pages 3813–3824, 2023. doi: 10.1109/ICCV51070.2023.00355.
- Hanqing Zhao, Dianmo Sheng, et al. X-Paste: Revisiting Scalable Copy-Paste for Instance Segmentation using CLIP and StableDiffusion. 2023.