Do Vision-Language Models Tackle Emptiness and Long-Tailed Distributions in Camera Trap Images?

Habiba Elkholy[1], **Imran Zualkernan**[1], **Jacky Judas**[2], **Roxanne Whelan**[3], **Andrew Gardner**[4], **Altaf Habib**[4]

[1]American University of Sharjah, [2]Soudah Development,
[3]King Abdullah University of Science and Technology, [4]Emirates Nature-WWF

Abstract

Camera traps allow low-cost and non-invasive monitoring of wildlife in natural habitats, allowing estimates of species diversity, abundance, and activity patterns. One of the biggest challenges in ecological AI is that camera trap data is inherently imbalanced. This imbalance is amplified by the inclusion of a large proportion of empty images, triggered by primitive and simple PIR sensors. This paper explores the use of Low-Rank Adaptation (LoRA) to fine-tune BioCLIP, a vision-language model (VLM), to address the problems of inherent and emptiness-induced imbalance. The approach is further extended by incorporating emptiness anchors to enhance BioCLIP's learning of emptiness. Experiments were conducted on two private datasets, as well as the Caltech Camera Traps (CCT), Wellington Camera Traps (WCT), iWildCam2022, Orinoquía Camera Traps (OCT), and Nkhotakota Camera Traps (NCT) datasets. Results show that as opposed to CNN-based transfer learning pipelines like InceptionV3 and two-stage approaches that incorporate Megadetector, the fine-tuned BioCLIP models perform better at identifying emptiness and minority classes. While in some cases the addition of the anchor leads to better overall accuracy, it has an adverse effect on rare classes. Further analysis of the learned embedding space is conducted to better understand the impact of the emptiness anchor.

1 Introduction

Due to several human-induced factors, such as habitat destruction and pollution, there is an ongoing decline in global biodiversity (Friese, 2020). One consequence is the increasing rate of species extinctions. In efforts to tackle this issue, ecologists employ non-invasive wildlife monitoring through camera traps. The captured data not only document the appearances of species, but can also provide insights into animal behaviors (O'Connell et al., 2011). Moreover, camera traps have proven to be instrumental in conservation efforts, enabling the re-discovery of species (Ahmed et al., 2016) or providing first-time sightings in previously undocumented regions (Al Hikmani et al., 2024). Despite the many benefits of camera traps, there is a prominent issue with the high amount of data that ecologists would have to manually review to identify present species. This process is time-consuming and labor-intensive, slowing critical conservation efforts. To address this bottleneck, researchers have increasingly turned to deep learning methods to automate species classification and accelerate data analysis. Nevertheless, developing efficient and accurate deep learning models remains challenging due to the prevalence of empty ("ghost") images. Such images do not include any animals, yet are captured when cameras are triggered by wind, light changes, people, or other non-target events (Granados et al., 2020). This exacerbates the existing class-imbalance issue in the domain of animal data, which naturally occurs due to the varying levels of species rarity.

A common approach to handle the imbalance caused specifically by empty images is to use a two-stage pipeline in which the first stage uses object detection and the latter uses classification, often using Convolutional Neural Networks (CNNs). However, this does not present an optimal solution, as a one-stage system is idealized. Other single-stage CNN-based approaches incorporate simple techniques such as oversampling to handle the imbalance, nonetheless finding that the data overfit (Zualkernan et al., 2022). Recent work (Rahmeh et al.) leveraged Low-Rank Adaptation to fine-tune BioCLIP, a vision-language model (VLM), on a private camera trap image dataset, showing that the approach can handle imbalance without the need for object detection or oversampling. This paper extends their approach to several camera trap datasets, including two private datasets, provided by Emirates Nature-WWF (used in prior work

(Rahmeh et al.)) and Soudah Development, as well as the Caltech Camera Traps (Beery et al., 2018), Wellington Camera Traps (Anton et al., 2018), iWildCam2022 (iWildCam) (Beuving et al., 2022), Orinoquía Camera Traps (Vélez et al., 2023), and Nkhotakota Camera Traps (Appel et al., 2025) datasets. Through this, we analyze whether this method generalizes, and how it deals with varying forms of imbalance. Furthermore, we experiment with the addition of an anchor representing the ghost class, analyzing its effect across the varying datasets.

2 Related Work

2.1 MegaDetector and SpeciesNet

Megadetector (MD) is an object detection model developed to detect people, vehicles, and animals in camera trap images (Beery et al., 2019). The YOLOv5x6 based MD model is often used in an ensemble with a classifier, such as Google's SpeciesNet (Gadot et al., 2024), which is based on the EfficientNet-V2-M architecture. Like MD, SpeciesNet was also trained on a large set of camera trap images, including blank images. Google's experiments show that both detection and cropping improve classification results. Firstly, MD runs on the image. If an object is detected, the image is cropped and inputted to the classifier, otherwise, it is labeled as blank. It is noteworthy that the training data used for both MD and SpeciesNet include four of the public datasets (CCT, WCT, iWildCam2022, and OCT) in their training data (Beery et al., 2019; Gadot et al., 2024). While their results are promising, there are two main issues with this method. Firstly, despite SpeciesNet being trained on blank images, the two-stage pipeline can still lead to propagating errors. Secondly, the CNN-based classifier is trained on a fixed number of classes, performing poorly on unseen species, highlighting its limitations under the domain shift. This motivates the exploration of VLMs.

2.2 Vision-Language Models

Unlike traditional CNNs which are designed solely for image classification, VLMs leverage a joint understanding of visual and textual information. VLMs learn a shared representation between images and corresponding descriptive raw text. OpenAI's Contrastive Language–Image Pretraining (CLIP) (Radford et al., 2021) is a popular VLM, which utilizes contrastive loss to maximize the similarity between matching image-text pairs, while minimizing the similarity between non-matching pairs. This makes it, like other VLMs, capable of zero-shot and few-shot generalization to unseen classes (Radford et al., 2021). Thus, VLMs are an attractive choice for the domain of wildlife monitoring, where imbalance and few-shot scenarios are prevalent. For example, Dussert et al. (Dussert et al., 2025) demonstrate CLIP's capability for zero-shot classification of animal behaviors in camera trap images. Additionally, (Tao et al., 2025) explore VLMs for remote sensing.

2.3 BioCLIP and LoRA

BioCLIP (Stevens et al., 2024) is a variant of CLIP that was specifically developed for biological classification. The model was trained on TreeOfLife-10M, a diverse dataset comprising millions of images of plants, animals, and fungi, each paired with its corresponding biological taxonomy, organized hierarchically as Kingdom, Phylum, Class, Order, Family, Genus, Species. While BioCLIP is trained on a large corpus of species, it has not been exposed to empty images, which are particularly relevant in the context of camera trap datasets, motivating fine-tuning on such data. However, BioCLIP (Stevens et al., 2024) has hundreds of millions of trainable parameters, which would make it very costly to fully fine-tune. To address this, (Rahmeh et al.) fine-tuned BioCLIP on a camera trap dataset using Low-Rank Adaptation (LoRA) (Hu et al., 2021), injecting low-rank matrices into the model's layers rather than updating all parameters. Their approach achieved a 94% accuracy and weighted F1-score, performing comparatively well on minority classes.

3 Methodology

3.1 Datasets

In this work, seven diverse camera trap image datasets are explored. All the included datasets include ghost images, and are deployed in varying regions, enabling the evaluation of BioCLIP under different domain conditions. The PD1 and PD2 datasets are proprietary datasets, while the remaining public datasets include the CCT (Beery et al., 2018), WCT (Anton et al., 2018), iWildCam (Beuving et al., 2022), OCT (Vélez et al., 2023), NCT (Appel et al., 2025) datasets. Notably, unlike most datasets, the WCT and OCT datasets do not have ghost images as their majority class; instead, the dominant classes are birds in WCT and collared

Dataset	No. Images	Original No. Classes	No. Mapped Classes	Percentage of "NA"	Imbalance Ratio	Shannon Entropy
PD1	66,352	15	15	52.02	417.20	0.5069
PD2	511,207	75	58	88.80	114394.25	0.1473
CCT	245,118	22	20	53.22	65231.00	0.6029
WCT	266,966	16	14	17.13	5723.15	0.5240
iWildCam	201,399	204	164	34.49	69469.00	0.5744
OCT	103,542	57	49	19.73	24784.00	0.6370
NCT	369,092	46	42	44.46	41028.25	0.5288

Table 1: Summary of datasets.

peccary in OCT. Given the nature of VLMs, the naming of the labels, i.e. the text paired with the image, is crucial. Since most datasets do not have species-level labels, the labels are standardized to indicate only the genus. In the case where the label refers to a higher taxonomic level, the label is mapped to the lowest applicable level. In this work, ghost images, labeled as "NA", refer to images that do not include an animal. This includes people (as in the PD2, OCT, and NCT datasets) and vehicles (in the CCT dataset), alongside blank images. Additionally, the images labeled as "unclassifiable" and "unknown" in the WCT (1.3%) and OCT (1.2%) datasets are excluded in all experiments.

Figure 1 shows the pairwise KL divergence between the distributions of the class counts in the datasets. Since each dataset has a different number of classes, the distribution of class counts is ordered from highest to lowest and interpolated to get a continuous distribution, allowing direct comparison. The KL divergence between two continuous distributions, P and Q, can be calculated as:

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, dx \qquad (1)$$

where $p(x)$ and $q(x)$ denote the probability density functions of x under distributions P and Q, respectively.

Notably, as seen in Figure 1, the PD2 dataset class distribution appears markedly different from all the other datasets, being closest to the iWildCam distribution. This vast difference is likely due to PD2 exhibiting a significantly higher percentage of ghosts. The iWildCam dataset appears to manifest the second most different class distribution, likely due to it having a relatively much higher number of classes.

Figure 1: The KL Divergence, in bits, between the dataset class distributions. The vertical axis represents distribution P, and the horizontal axis represents distribution Q. A value of 0 indicates identical distributions, while higher values indicate a larger difference between the distributions.

3.2 Baselines

Two baselines are run to compare to our approach. First, transfer learning using InceptionV3 is leveraged, where the pretrained backbone is frozen and only the classification layers are trained. Next, we run the MD and SpeciesNet ensemble (Gadot et al., 2024) described in Section 2.1 on all datasets. SpeciesNet was trained using specific class names based on the taxonomy. However, since SpeciesNet is a CNN and not a VLM, it will not be able to predict a class name that it was not trained on. Therefore, we refer to the taxonomy mapping which is released by the Labelled Information Library of Alexandria: Biology and Conservation (LILA BC) for all the included public datasets (LILA, 2026) to map class names before running the ensemble.

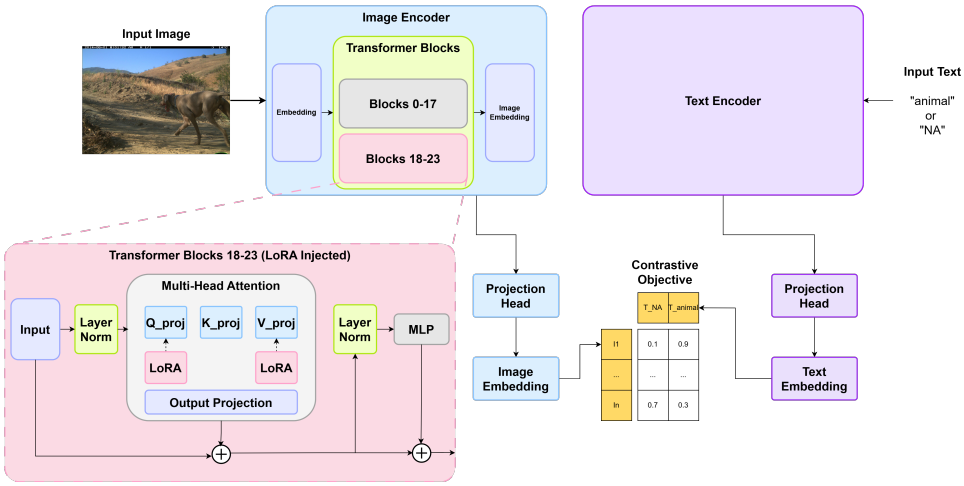


Figure 2: Fine-tuning BioCLIP with LoRA.

3.3 Fine-tuning BioCLIP with LoRA

Following previous work (Rahmeh et al.), we fine-tune BioCLIP using the LoRA technique. As with standard practice, the text encoder is frozen. For the image encoder, which consists of a vision transformer (Stevens et al., 2024), LoRA adapters are injected into the last 6 residual blocks only. These are the high-level blocks, allowing the fine-tuned model to adapt to the high-level semantics pertaining to camera trap images, with minimal disruption to pretrained representations. The low-level blocks, i.e. the first 6 blocks, encode low-level features such as edges, color, and texture. These features are stable across domains, therefore modifying them using our dataset, which is significantly smaller than TreeOfLife-10M (Stevens et al., 2024), would disrupt pretrained invariances, while increasing the risk of overfitting. Blocks 7 to 12 represent the mid-level blocks that encode shapes, poses, and rough object layouts. While modifying them may handle subtle domain shifts, adapting them would increase training cost and optimization instability. Accordingly, LoRA adapters are injected into the query and value projection matrices of the self-attention layers in the selected transformer blocks, while the key projections are kept frozen. Adjusting the query projections allows rare-class tokens to form new attention patterns, learning visual features uncommonly seen during pretraining. As well as that, adjusting value projections allows attended regions to have distinct embedding signatures, even if the visual evidence is subtle. On the other hand, keeping the key projections frozen allows major classes to remain unaffected, as the overall context remains stable. Figure 2 illustrates the architecture of fine-tuning BioCLIP with LoRA.

3.4 Ghost Anchor

Our method is modified further by adding a ghost image anchor. Essentially, for each image in the training set, a ghost image is randomly selected. During training, the cosine similarity between the query image embedding and its corresponding ghost anchor image embedding is maximized if they are both ghosts, and minimized if the query is not a ghost image. This modification aims to enhance BioCLIP's learning of ghosts, as the base model is not trained on emptiness. The cosine similarity between the query image embedding, q, and the anchor image embedding, a, is calculated as:

$$\text{sim}(\mathbf{q}, \mathbf{a}) = \frac{\mathbf{q} \cdot \mathbf{a}}{\|\mathbf{q}\| \, \|\mathbf{a}\|} \tag{2}$$

$\text{sim}(\mathbf{q}, \mathbf{a}) \in [-1, 1]$, where a value of 1 denotes that the vectors are identical in direction, indicating similarity. A value of 0 indicates the two vectors are orthogonal, while a value of -1 indicates that the vectors are in opposite directions. Thus, the distance between the query and anchor image embeddings is denoted as:

$$d(\mathbf{q}, \mathbf{a}) = 1 - \text{sim}(\mathbf{q}, \mathbf{a}) \tag{3}$$

This is used to calculate the contrastive loss term, using a flag, y. $y = 1$ if the query is also a ghost like the anchor, and $y = 0$ if the query is a non-ghost:

$$\mathcal{L}(\mathbf{q}, \mathbf{a}, y) = y \, d(\mathbf{q}, \mathbf{a})^2 + (1 - y) \, \max(0, m - d(\mathbf{q}, \mathbf{a}))^2 \tag{4}$$

Dataset	InceptionV3	MD + SpeciesNet	Zero-Shot BioCLIP	Fine-Tuned BioCLIP	Fine-Tuned BioCLIP + Image Anchor
PD1	0.93 / 0.41	0.77 / 0.40	0.21 / 0.28	**0.96 / 0.87**	**0.96** / 0.86
PD2	0.93 / 0.37	0.85 / 0.12	0.008 / 0.044	**0.98 / 0.72**	**0.98** / 0.67
CCT	0.71 / 0.35	0.86 / 0.67	0.19 / 0.18	0.42 / **0.74**	**0.94** / 0.73
WCT	0.83 / 0.50	0.85 / **0.89**	0.05 / 0.08	0.86 / 0.78	**0.87** / 0.78
iWildCam	0.69 / 0.36	0.67 / 0.43	0.06 / 0.03	0.90 / **0.61**	**0.91** / 0.47
OCT	0.48 / 0.16	0.50 / 0.35	0.02 / 0.02	**0.86 / 0.57**	**0.86** / 0.55
NCT	0.64 / 0.17	0.69 / **0.75**	0.07 / 0.03	**0.93** / 0.69	**0.93** / 0.69

Table 2: Performance of models across datasets (Accuracy / Macro-Avg F1).

Dataset	InceptionV3	MD + SpeciesNet	Zero-Shot BioCLIP	Fine-Tuned BioCLIP	Fine-Tuned BioCLIP + Image Anchor
PD1	0.88 / 0.40	0.90 / 0.26	0.00 / 0.12	**0.97 / 0.78**	0.97 / 0.71
PD2	0.97 / 0.29	0.94 / 0.00	0.00 / 0.00	**0.99** / 0.48	**0.99 / 0.62**
CCT	0.86 / 0.03	0.92 / 0.32	0.00 / 0.0025	0.00 / **0.55**	**0.97** / 0.27
WCT	0.10 / 0.035	**0.80** / 0.39	0.00 / 0.00	0.63 / **0.47**	0.63 / 0.42
iWildCam	0.88 / 0.00	0.84 / **0.40**	0.00 / 0.00	0.97 / 0.34	**0.98** / 0.14
OCT	0.49 / 0.00	0.65 / **0.17**	0.00 / 0.00	**0.85** / 0.00	**0.85** / 0.00
NCT	0.80 / 0.13	0.97 / 0.22	0.00 / 0.00	0.97 / **0.40**	**0.98** / 0.37

Table 3: Ghost F1-Score / 20% Rarest Classes Macro-Avg F1-Score across models and datasets.

where m denotes the margin hyperparameter, which sets a threshold distance for dissimilar pairs. Our experiments use $m = 0.5$. Furthermore, the loss term, $\mathcal{L}(\mathbf{q}, \mathbf{a}, y)$, is added to BioCLIP's original loss term, with no weighting.

3.5 Experimental Setup

For all experiments that involve fine-tuning BioCLIP with LoRA, the images were resized to 224x224 pixels, with the LoRA rank set to 8 and the scaling factor set to 16. Furthermore, data augmentation was applied to the training data, including random rotations (up to 30°), horizontal and vertical flips, color jittering, and affine transformations to simulate real-world variation. However, only resizing and normalization were applied to the validation and test data to ensure consistency. The models were trained using the AdamW optimizer with a base learning rate of $1\mathrm{e}{-5}$, a weight decay of $1\mathrm{e}{-4}$, and a cosine learning rate schedule with linear warm-up. All experiments followed a 70-15-15 train-validation-test split and were trained for 15 epochs with a batch size of 128, where the model with the best validation accuracy was selected.

Model	Rare Classes F1
InceptionV3	0.00
MD + SpeciesNet	**0.33**
Zero-Shot BioCLIP	0.00
Fine-Tuned BioCLIP	0.13
Fine-Tuned BioCLIP + Image Anchor	0.13

Table 4: Macro-Average F1-score of the 30% rarest classes in the OCT dataset.

4 Results and Analysis

Table 2 summarizes the accuracy and macro-average F1-scores across all models and datasets. Additionally, table 3 summarizes the F1-score of the ghost class and the macro-average F1-score of the 20% rarest classes.

Main findings: Zero-shot BioCLIP consistently underperformed, achieving a ghost F1-score of zero across all datasets, rendering it incapable of recognizing ghosts. This is a consequence of the base BioCLIP model not being trained on any ghost images, thereby lacking its representation. In general, fine-tuning BioCLIP leads to better results

(a) An image of a bat.

(b) An image of an insect.

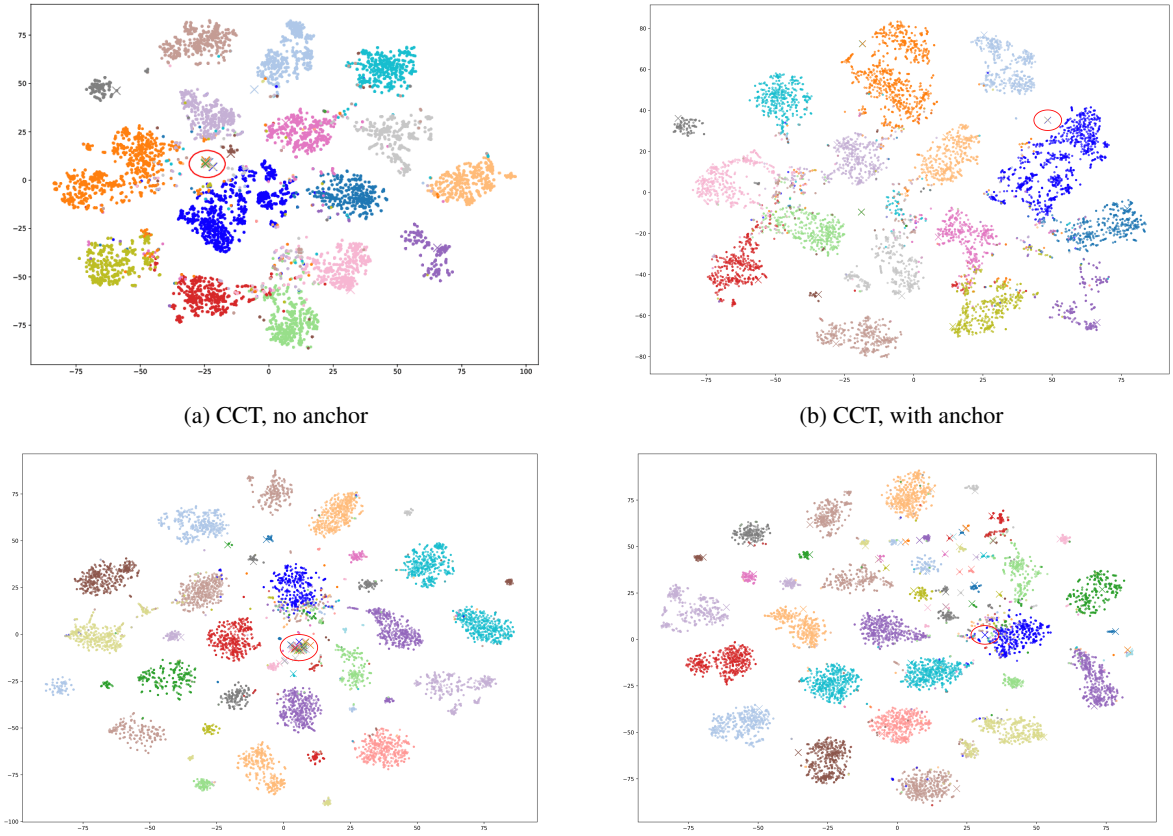
Figure 3: Examples from the CCT dataset.

compared to the two CNN-based methods. Additionally, while adding the ghost anchor during fine-tuning of BioCLIP results in only minor changes in accuracy for most datasets, its impact is notably significant for the CCT dataset. Finally, apart from the PD2 dataset, the ghost anchor degrades performance on the rare classes, while generally only increasing the ghost classification slightly.

BioCLIP and ghosts in the CCT dataset: Unlike with the other datasets, for the CCT dataset, fine-tuning BioCLIP performs well only with the addition of the ghost anchor. With the addition of the anchor, the accuracy increases significantly from 42% to 94%. Otherwise, the model confuses the ghost images with the rarest classes; either bats (one instance in the training set) or insects (two training instances). Therefore, the ghost anchor is necessary in this case. Figures 3a and 3b show examples of the bat and insect images, respectively, from the CCT dataset. Visually, the images appear nearly empty, providing further insight into why the models confuse the ghost class with them.

Ghost anchor effect on minority classes: Overall, adding the anchor worsens the macro-average F1-score for the rarest 20% classes, except for PD2. This may be due to the anchor being specific to ghosts, which is mostly the dominant class. As a result, the training loss may place disproportionate emphasis on the ghost class, causing the model to underperform on rare classes. Nevertheless, for the PD2 dataset, which has the most different class distribution as illustrated by Figure 1, adding the anchor yields a 14% higher macro-average F1-score for the rare classes, despite that approach achieving a lower overall macro-average F1-score.

Comparison to CNN-based baselines: In general, the MD and SpeciesNet ensemble performs better than transfer learning using InceptionV3 on the public datasets. However, the latter outperforms the ensemble on the private datasets, likely as they were not included in the training data of the MD and SpeciesNet ensemble, highlighting its struggle to generalize to new data. In general, the fine-tuned BioCLIP models outperform the CNN-based models, with a few exceptions. For instance, for the WCT dataset, the MD and SpeciesNet ensemble outperforms the fine-tuned BioCLIP models in terms of ghost classification. This exception may be due to the fact that the WCT dataset has the lowest percentage of ghosts, specifically 17.13%, whereas the major class, birds, comprises 57.88% of the dataset, indicating the ensemble may struggle when the ghost images are more dominant. In addition, while the ensemble yields higher overall macro-average F1 score (by a significant 11%), the fine-tuned BioCLIP models yield better performance on the rare classes. This may be because the WCT dataset, which has only 14 classes, has mostly common classes (e.g. cat, dog, deer), which SpeciesNet was frequently exposed to during training. Thus, while it can perform well on such dominant classes, SpeciesNet struggles with rare species. This is also seen in the NCT dataset, where the ensemble yields a gain of 6% in overall macro-average F1 score, but a loss of 18% in the rare classes macro-average F1-score, which may be due to the two datasets exhibiting similar normalized Shannon entropy scores, as seen in Table 1, indicating similar class imbalance. In general, all models struggle with the rare classes in the OCT dataset. Table 2 shows that all models cannot classify the rarest 20% classes, apart from the MD and SpeciesNet ensemble. However, this macro-average of 17% is achieved by only one sample across all the rarest classes. This is due to the rarest classes having only very few samples, for which some only appear during training and not testing, and vice versa. Therefore, for the OCT dataset,



(a) CCT, no anchor

(b) CCT, with anchor

(c) PD2, no anchor

(d) PD2, with anchor

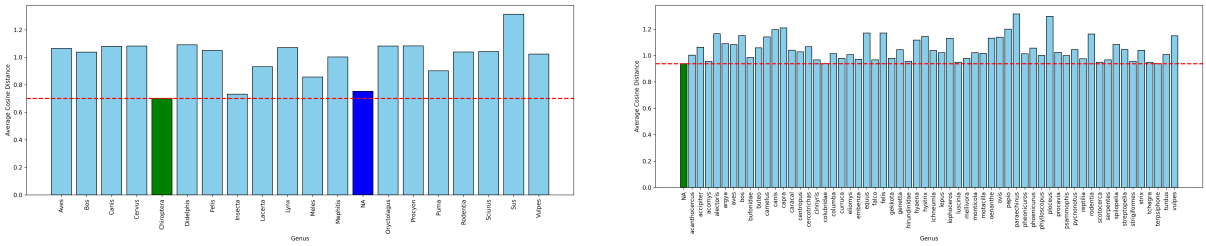
Figure 4: t-SNE visualizations of image and text embeddings encoded by fine-tuned BioCLIP across datasets and training setups. Crosses denote text embeddings and dots denote image embeddings. The dark blue color represents the ghost class, and the circled regions contain the ghost text embedding.

the rarest 20% do not provide enough insights. Table 4 shows the macro-average F1-scores on the rarest 30% classes. While the fine-tuned BioCLIP models perform better than InceptionV3 and zero-shot BioCLIP, the MD and SpeciesNet ensemble outperforms all.

4.1 Latent Space Analysis

Latent space analysis was conducted by first using the fine-tuned BioCLIP encoders to extract the text and image embeddings and then using t-SNE to visualize them, providing insights into the distances between image and text embeddings. Since the datasets are large, subsamples were used for the t-SNE visualization. Note that Figures 4c and 4d, which correspond to the t-SNE plots of the PD2 embeddings, do not have visually distinct colors for all classes, as there are too many classes. However, the color representing the ghosts is unique (dark blue).

As noted earlier, fine-tuning BioCLIP on the CCT dataset only performs well with the addition of the ghost anchor. Figure 4a shows the latent space of the image and text embeddings obtained from the encoders of the fine-tuned BioCLIP model, without the ghost anchor. An annotated cluster of text embeddings is visible, which includes the text embeddings corresponding to the bat, insect, and ghost. For further analysis, the average cosine distance between the image embeddings of each class and all corresponding text embeddings is computed and plotted as seen in Figure 5a. On average, the ghost image embeddings are closer to the bat text embedding, followed by the insect text embedding, than the "NA" text embedding. The difference between the average distance from the "NA" text embedding is only 0.02 and 0.05 units, compared to the bat and insect text embeddings, respectively. Figure 4b shows the latent space after adding the ghost anchor, in which the text embeddings are more clearly separated from one another. Thus, in this case, adding the ghost anchor allows the model to better differentiate between ghosts and bats and insects, as the ghost



(a) CCT

(b) PD2

Figure 5: Average cosine distances between ghost image embeddings and all text embeddings for models fine-tuned without an anchor. Green bars denote the minimum average distance. The blue bar in (a) corresponds to the "NA" text embedding, and in (b), this value is the minimum.

Dataset	Image Anchor	Text Anchor
PD1	0.96 / 0.86	0.96 / 0.85
PD2	0.98 / 0.67	0.98 / 0.67
CCT	0.94 / 0.73	0.94 / 0.73

Table 5: Results using a text anchor rather than an image anchor (Acc / Macro-Avg F1).

images are pushed further from the bat and insect images.

While both the models with and without the image anchor are able to distinguish the ghosts in PD2, adding the anchor improves the inter-class separation of the image and text embeddings. Figure 4c shows groups of text embeddings clustered together. However, as seen in Figure 5b, the "NA" text embedding is the closest to the ghost image embeddings, on average, allowing the model to perform well on ghosts. Adding the anchor diminishes this grouping of text embeddings, as seen in Figure 4d. Text embeddings appear to be close to their corresponding image embeddings, even for minority classes, improving the macro-average F1-score on rare classes.

4.2 Ablation Study: Text Anchor

In the earlier experiments, our anchor consisted of a ghost image. Here, we instead use a text-based ghost anchor rather than an image. The added contrastive loss encourages the model to focus on image–text similarity, as in CLIP and BioCLIP, rather than on direct image–image similarity. The two anchor variants are theoretically equivalent in BioCLIP's shared image-text embedding space: whether the emptiness anchor is represented as an image embedding or the fixed "NA" text embedding, non-ghost samples are pushed away from the ghost representation, inducing the same geometric separation. This equivalence is demonstrated empirically by these experiments. Additionally, since the text embeddings are fixed, they do not need to be recomputed at every training step, unlike the image embeddings. Moreover, the model no longer needs to store the added anchor images, making the training process both computationally and memory efficient. Table 5 compares the results of using the image versus text ghost anchor for the two private datasets as well as the CCT dataset. There is no significant difference between the two approaches, making the text anchor the more efficient option.

5 Conclusions

This paper explores how fine-tuning BioCLIP with LoRA on camera trap datasets addresses ghost images and class imbalance. Experiments are conducted on seven camera trap datasets from different regions and with varying levels of imbalance, showing that in general, fine-tuning BioCLIP with LoRA can handle ghosts and minority classes without the need for oversampling or MegaDetector. In general, with a few exceptions, the fine-tuned BioCLIP models outperform InceptionV3 and the two-stage MD and SpeciesNet approach, showing better overall accuracy and generalization to rare genera. Modification through adding a ghost anchor image significantly improves performance on the CCT dataset. Nevertheless, across most datasets, the addition of the anchor degrades performance on rare classes, as more emphasis is placed on the majority ghost class. Thus, since the ghost anchor is necessary for datasets like CCT, better integration strategies are needed. Overall, these results demonstrate that VLM-based fine-tuning is a promising direction for camera trap species classification, particularly in imbalanced settings.

Limitations

One limitation of our approach is that it operates only at the genus level, without leveraging the full taxonomy available in BioCLIP. Additionally, since the text anchor is equally performant but more computationally efficient than the image anchor, future work should repeat the full experiments in this setting. Furthermore, while the ghost anchor improves ghost classification, it degrades minority class performance. Therefore, future work will explore extending the anchor concept to rare classes, as well as warming up the model with the anchor in early epochs and removing it later to mitigate overemphasis on the ghost class. Modifications to the loss term to improve cross-dataset generalization will also be investigated, alongside cross-dataset evaluation of trained models to assess transferability across domains and deployment conditions.

Acknowledgements

This work was supported by the American University of Sharjah Faculty Research Grant (FRG-25C-E22).

References

Shakeel Ahmed, Rashid Al Zaabi, Pritpal Soorae, Junid N. Shah, Eissa Al Hammadi, Ricardo Pusey, and Sheikha Al Dhaheri. 2016. Rediscovering the Arabian sand cat (Felis margarita harrisoni) after a gap of 10 years using camera traps in the Western Region of Abu Dhabi, United Arab Emirates. *European Journal of Wildlife Research*, 62(5):627–631.

Hadi Al Hikmani, Khalid Al Hikmani, and Abdullah Alshehri. 2024. First camera-trap record of caracal twin kittens in Saudi Arabia. *Oryx*, 58(1):14–14.

Victor Anton, Stephen Hartley, Andre Geldenhuis, and Heiko U Wittmer. 2018. Monitoring the mammalian fauna of urban areas using remote cameras and citizen science. *Journal of Urban Ecology*, 4(1).

Cara L. Appel, Ashwin Subramanian, Jonathan S. Koning, Marnet Ngosi, Christopher M. Sullivan, Taal Levi, and Damon B. Lesmeister. 2025. Developing custom computer vision models with njobvu-ai: A collaborative, user-friendly platform for ecological research. *Ecological Applications*, 35(6):e70096.

Sara Beery, Dan Morris, and Siyu Yang. 2019. Efficient Pipeline for Camera Trap Image Review. *arXiv preprint*.

Sara Beery, Grant Van Horn, and Pietro Perona. 2018. Recognition in Terra Incognita. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11220, pages 472–489. Springer International Publishing, Cham.

John Beuving, Phil Culliton, Sara Beery, and Ștefan Istrate. 2022. iwildcam 2022 - fgvc9. Kaggle competition.

Gaspard Dussert, Vincent Miele, Colin Van Reeth, Anne Delestrade, Stéphane Dray, and Simon Chamaillé-Jammes. 2025. Zero-shot animal behaviour classification with vision-language foundation models. *Methods in Ecology and Evolution*, 16(7):1460–1472.

Carrie Friese. 2020. *Cloning Wild Life: Zoos, Captivity, and the Future of Endangered Animals*. New York University Press.

Tomer Gadot, Ștefan Istrate, Hyungwon Kim, Dan Morris, Sara Beery, Tanya Birch, and Jorge Ahumada. 2024. To crop or not to crop: Comparing whole-image and cropped classification on a large dataset of camera trap images. *IET Computer Vision*, 18(8):1193–1208.

Joseph Granados, Chris Halle, and Gurman Gill. 2020. Classifying False Alarms in Camera Trap Images using Convolutional Neural Networks. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1445–1451, Las Vegas, NV, USA. IEEE.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint*.

LILA. 2026. Taxonomy mapping for camera trap datasets. https://lila.science/taxonomy-mapping-for-camera-trap-data-sets/. Labeled Information Library of Alexandria: Biology and Conservation. Accessed January 19, 2026.

Allan F. O'Connell, James D. Nichols, and K. Ullas Karanth, editors. 2011. *Camera Traps in Animal Ecology*. Springer Japan, Tokyo.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.

Nesreen Rahmeh, Maryam Gharaibeh, Imran Zualkernan, Andrew Gardner, and Altaf Habib. Fine-tuning vision-language models enables robust classification of naturally imbalanced camera trap images. In *Eighteenth International Conference on Machine Vision (ICMV 2025)*, page 102. SPIE.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. 2024. BioCLIP: A Vision Foundation Model for the Tree of Life. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19412–19424, Seattle, WA, USA. IEEE.

Lijie Tao, Haokui Zhang, Haizhao Jing, Yu Liu, Dawei Yan, Guoting Wei, and Xizhe Xue. 2025. Advancements in Vision–Language Models for Remote Sensing: Datasets, Capabilities, and Enhancement Techniques. *Remote Sensing*, 17(1):162.

Juliana Vélez, William McShea, Hila Shamon, Paula J. Castiblanco-Camacho, Michael A. Tabak, Carl Chalmers, Paul Fergus, and John Fieberg. 2023. An evaluation of platforms for processing camera-trap data using artificial intelligence. *Methods in Ecology and Evolution*, 14(2):459–477.

Imran Zualkernan, Salam Dhou, Jacky Judas, Ali Reza Sajun, Brylle Ryan Gomez, and Lana Alhaj Hussain. 2022. An IoT System Using Deep Learning to Classify Camera Trap Images on the Edge. *Computers*, 11(1):13.