Wildlife Classification on Video from Phototraps in the Alta Murgia Park.

Vito Proscia ✉
Dipartimento di informatica, Uniba, Bari

Abstract: Local wildlife monitoring in the Alta Murgia region is essential for both species conservation and environmental protection. This paper addresses the problem of automatic animal recognition in the area, with the aim of supporting naturalists involved in wildlife monitoring. We propose a complete pipeline, starting from video data and leading to the classification of the detected animals. The study also tackles the issue of class imbalance. In particular, we trained both traditional classifiers and fine-tuned the VitDet model, achieving an accuracy above 95%, demonstrating robust classification performance across animal classes.

1 Introduction

Wildlife recognition in the Alta Murgia region is a fundamental requirement for ecological research and conservation planning. In particular, monitoring of the local fauna is carried out by collecting video footage from camera traps installed at strategic points throughout the national park. This monitoring supports the study of various ecological factors, such as the number of species present, their behaviors, relative abundances, and both intra- and interspecific interactions. Traditionally, the collected videos were analyzed and annotated manually based on the species observed in the footage. While effective, this approach is highly time-consuming and limits the scalability of monitoring efforts. To improve the efficiency of this process, we propose a pipeline that automates species recognition directly from video data, addressing class imbalance issues and evaluating different classification models trained using two strategies: fine-tuning of VitDet [2] and feature extraction for traditional classifiers. The experimental results demonstrate that the proposed approach can significantly support wildlife monitoring, enhancing conservation efforts and providing valuable tools for ecologists and park managers.

2 Related Work

The automatic recognition of wildlife from camera trap data has gained increasing attention due to its potential to reduce manual labor and enhance monitoring efficiency. Several approaches have been proposed, with a particular focus on transfer learning techniques. For instance, [4] employed the Xception model pre-trained on Keras, adding convolutional layers to its output. This model was trained on a high-resolution wildlife image dataset, in contrast to our setting, which involves noisy video data.

Another notable contribution comes from a study conducted in the Netherlands [3], aimed at supporting the conservation of declining grassland bird species. The authors evaluated the performance of various Vision Transformer models, including DETA and OWLVIT, using images collected in natural environments. Their study explicitly distinguished between day and night conditions and incorporated a video processing pipeline that included MegaDetector [1] for animal detection. In contrast, our approach processes both day and night images together within a unified framework.

3 Data

3.1 Data description

In order to train and evaluate our wildlife recognition pipeline, we relied on a custom dataset composed of videos captured by camera traps deployed in the natural environments of the Alta Murgia National Park (Italy). The footage documents various species of local fauna and includes both diurnal (RGB) and nocturnal (infrared, wavelength = 850 nm) recordings. This dual condition introduces significant variability and presents several challenges, such as motion blur, background clutter (e.g., vegetation movement), varying lighting conditions, and reduced visibility—especially in low-light or night-time scenarios. The dataset consists of over 700 video clips, each ranging in duration from approximately 20 to 30 seconds. These videos were acquired in real-world monitoring settings, often triggered by animal movement, which adds to the ecological validity of the data but also increases the presence of noise and false positives (e.g., vegetation or lighting changes). All video sequences were manually reviewed and annotated for the presence and identity of animal species, forming a reliable ground truth for training and evaluation purposes. The diversity in environmental conditions, species, and recording quality

makes this dataset particularly suitable for testing the robustness of wildlife detection models in unconstrained outdoor scenarios.

Figure 1: RGB frame captured during daylight.

Figure 2: Infrared frame captured at night.

The dataset contains instances from a total of 15 animal classes, with a highly imbalanced distribution that reflects the natural occurrence frequency of the species in the park. The class distribution is described in Table 1.

It is important to note that some classes in the dataset (e.g., Bug, Mouse, and Snake) have zero instances. These classes were initially included based on expert recommendations from domain specialists familiar with the fauna of the Alta Murgia region. However, no corresponding occurrences were identified during the manual annotation phase, likely due to the specific behavior, size, or low detectability of these species within the camera trap footage collected. This absence of instances for certain categories led to their exclusion from the training phase. As a result, performance evaluation is limited to the classes present in the annotated data. Nonetheless, these classes have been retained in the label map to maintain consistency with the original taxonomy and to enable future integration of additional instances as new data becomes available.

Class	Number of Instances
Podolic cow	304
Bird	200
Boar	141
Fox	61
Wolf	21
Cat	12
Dog	3
Badger	4
Porcupine	4
Lizard	4
Butterfly	1
Weasel	1
Bug	0
Mouse	0
Snake	0

Table 1: Class distribution of annotated wildlife instances in the dataset.

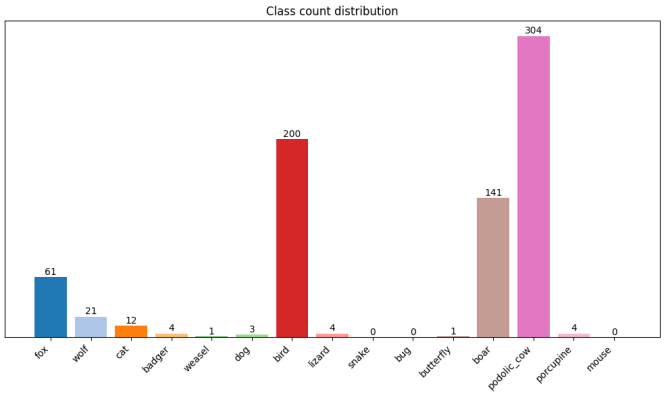

Figure 3: Bar plot showing the frequency of each class in the dataset.

3.2 Dataset Construction

To build the dataset used in our experiments, we designed a data processing pipeline that transforms raw video footage into a structured set of annotated frames containing wildlife instances. The main steps of this pipeline are as follows:

1. **Video Collection**:
   Over 700 videos were collected using fixed-position camera traps installed in various locations across the Alta Murgia National Park. The recordings include both daytime (RGB) and nighttime (infrared at 850 nm) footage, with durations ranging from 20 to 30 seconds per video.

2. **Frame Sampling**:
   All videos were processed using MegaDetector [1] as an object detector. Each video was sampled every 5 frames, and the frame with the highest confidence score for the `animal` label was selected. For each selected frame, we stored the image, the bounding box coordinates in COCO format (x, y,

width, height), and the species label, which was extracted from the video filename.

3. **Data Augmentation**:
   One of the major challenges encountered was the class imbalance, due to the dominance of certain animal species in the dataset. To address this, we artificially increased the number of samples for underrepresented classes by applying a series of transformations to the cropped animal frames based on the bounding boxes. The following augmentations were applied:

   - **Random Horizontal Flip**: Simulates mirrored scenarios to enhance the model's robustness to orientation changes.
   - **Random Rotation (±20°)**: Introduces angular variability to account for slight camera tilt or different animal poses.
   - **Color Jittering (±0.2 brightness, contrast, saturation)**: Alters lighting and color conditions to simulate variations in exposure, particularly beneficial for RGB frames.

   These transformations were implemented using PyTorch's `transforms.Compose()` and applied exclusively to minority class samples during training, up to a minimum of 75 frames per underrepresented class, in order to avoid excessive synthetic data.

At the end of this process, we obtained a total of 1321 annotated images (starting from 757 original ones).

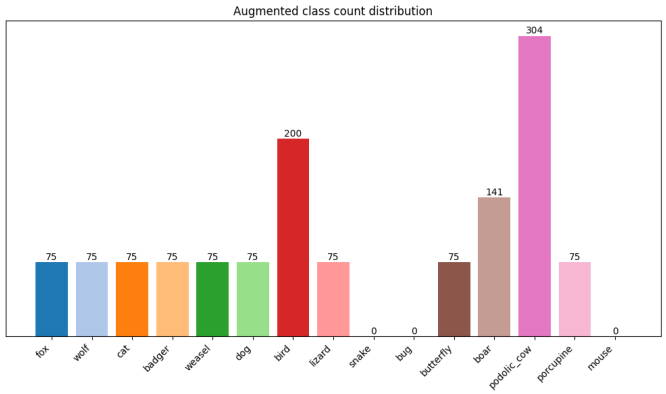

Figure 4: Bar plot showing the frequency of each class in the augmented dataset.

4 Methods

To achieve our goal of classifying local wildlife species, we explored two complementary approaches, both based on the VitDet architecture [2], but following different strategies.
Before feeding the images into the VitDet model for fine-tuning, a specific preprocessing pipeline was applied to ensure compatibility with the model's input requirements. The transformation included resizing all images to 224×224 pixels, converting them to PyTorch tensors, and normalizing the pixel values with mean and standard deviation of $(0.5, 0.5, 0.5)$.

4.1 Vision Transformer Fine-Tuning

We fine-tuned the `vit_base_patch16_224` model. Initially, the backbone was frozen to train only the classification head, and subsequently unfrozen to allow full end-to-end training. The classification head was adapted to match the number of species in our dataset. In both cases we trained the models for 5 epochs and we used the Adam optimizer with a learning rate of 1×10^{-3}. The loss function used was the standard cross-entropy loss, suitable for multi-class classification tasks.

4.2 Hybrid Approach

We used VitDet as a fixed feature extractor to generate visual embeddings from input images. The feature vectors, of size 768, were extracted from the final transformer block before the classification head. These embeddings capture high-level visual semantics of the input frame. We then trained traditional machine learning models, including Linear Regression, Decision Tree, and K-Nearest Neighbors, using the scikit-learn library.

All models were trained using an NVIDIA RTX 4070 GPU (12 GB VRAM).

5 Experiments and Results

To assess the performance of the proposed approaches, we conducted a series of experiments using the labeled dataset described in Section 3. Each model was evaluated using stratified 5-fold cross-validation, ensuring that the class distribution of species was preserved across all training and test splits. We report the average results across folds using standard classification metrics: accuracy, precision, recall, and F1-score. Experiments were carried out independently for both the fine-tuned Vision Transformer and the hybrid approach based on visual feature extraction followed by training of classical classifiers. Analysis of the obtained results reveals that the best-performing model was the hybrid approach with Logistic Regression, followed closely by the K-Nearest Neighbors (KNN) classifier. These results highlight the strong discriminative power of the features extracted by VitDet, which, when combined with robust classifiers, yield highly accurate predictions. The use of compact, meaningful representations helped these models generalize well, achieving high performance with relatively low variability across folds. The Vision Transformer model fine-tuned with a frozen backbone also achieved competitive performance; however, its standard deviation was slightly higher, suggesting greater

variability in its predictions across different cross-validation splits. This may indicate a certain sensitivity to specific data partitions. On the other hand, the models that performed worst were the Decision Tree and the fully fine-tuned ViT model (with all weights trainable). These approaches showed not only lower average performance across all metrics but also significantly higher variance, indicating instability and limited generalization capacity under the current dataset conditions.
A detailed summary of all evaluation metrics is presented in 6, which provides a comprehensive comparison of the models across the four main performance indicators.

6 Conclusions

This study investigated effective solutions for classifying wildlife species from challenging video data collected in natural environments. Two main approaches were explored: (i) fine-tuning a Vision Transformer (ViT) model, both with a frozen backbone and allowing full weight updates, and (ii) a hybrid strategy where ViT was used as a feature extractor and the resulting embeddings were used to train classical machine learning classifiers such as Logistic Regression, Decision Tree, and K-Nearest Neighbors. After a detailed analysis based on stratified 5-fold cross-validation, the hybrid approach combined with Logistic Regression emerged as the most effective model in distinguishing between various animal species, achieving performance metrics exceeding 95% across all evaluation criteria. However, the lack of annotated examples for some underrepresented classes suggests that collecting more instances for those species (e.g., bug, snake, and mouse) could improve the overall classification performance and robustness of the models. These classes were retained in the label map to accommodate future expansions of the dataset. Future work could focus on enhancing video frames through image enhancement techniques in the data preprocessing pipeline to further improve classification accuracy. Additionally, incorporating contextual or temporal information beyond static images might provide further gains. To make the results actionable for field biologists and conservationists, a dedicated software interface should be developed to allow efficient interaction with the trained models and facilitate real-world usage.

References

[1] Mitchell Fennell, Christopher Beirne **and** A. Cole Burton. **?**Use of object detection in camera trap image identification: Assessing a method to rapidly and accurately classify human and animal detections for research and application in recreation ecology**?** **in***Global Ecology and Conservation*: 35 (2022), e02104. ISSN: 2351-9894. DOI: 10.1016/j.gecco.2022.e02104. URL: https://www.sciencedirect.com/science/article/pii/S2351989422001068.

[2] Yanghao Li **andothers**. *Exploring Plain Vision Transformer Backbones for Object Detection*. 2022. arXiv: 2203.16527 [cs.CV]. URL: https://arxiv.org/abs/2203.16527.

[3] L.M. Trusca. *Investigating suitable vision transformer models for wildlife camera trap data*. **february** 2025. URL: http://essay.utwente.nl/105228/.

[4] Xihao Wang, Peihan Li **and** Chengxi Zhu. **?**Classification of Wildlife Based on Transfer Learning**?** **in***Proceedings of the 2020 4th International Conference on Video and Image Processing*: ICVIP '20. Xi'an, China: Association for Computing Machinery, 2021, **pages** 236–240. ISBN: 9781450389075. DOI: 10.1145/3447450.3447487. URL: https://doi.org/10.1145/3447450.3447487.

Tables

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.9811 ± 0.0115	0.9779 ± 0.0462	0.9823 ± 0.0441	0.9793 ± 0.0123
KNN (k=1)	0.9705 ± 0.0147	0.9635 ± 0.0584	0.9728 ± 0.0681	0.9665 ± 0.0533
VitDet (frozen)	0.9492 ± 0.0114	0.9390 ± 0.0876	0.9367 ± 0.1087	0.9322 ± 0.0177
Decision Tree	0.8303 ± 0.0138	0.8072 ± 0.1314	0.8056 ± 0.1538	0.8015 ± 0.1334
VitDet (full FT)	0.4614 ± 0.1222	0.2852 ± 0.3199	0.3490 ± 0.3897	0.2811 ± 0.3043

Table 1: Performance comparison of all models (mean ± std across 5 folds)

True \ Pred	Badger	Bird	Boar	Butterfly	Cat	Dog	Fox	Lizard	Cow	Porcupine	Weasel	Wolf
Badger	15.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Bird	0.0 ± 0.0	39.2 ± 1.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.2 ± 0.4	0.4 ± 0.5	0.2 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Boar	0.2 ± 0.4	0.0 ± 0.0	27.0 ± 0.9	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.2 ± 0.4	0.2 ± 0.4	0.2 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	0.4 ± 0.5
Butterfly	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	15.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Cat	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	15.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Dog	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	15.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Fox	0.0 ± 0.0	0.4 ± 0.8	0.0 ± 0.0	0.0 ± 0.0	0.2 ± 0.4	0.0 ± 0.0	13.2 ± 1.3	0.2 ± 0.4	0.0 ± 0.0	0.2 ± 0.4	0.0 ± 0.0	0.8 ± 1.2
Lizard	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	15.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Cow	0.0 ± 0.0	0.0 ± 0.0	0.6 ± 0.8	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	59.8 ± 1.0	0.0 ± 0.0	0.0 ± 0.0	0.4 ± 0.5
Porcupine	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	15.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Weasel	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	15.0 ± 0.0	0.0 ± 0.0
Wolf	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.2 ± 0.4	0.0 ± 0.0	14.8 ± 0.4

Table 2: Confusion matrix of Logistic Regression (mean ± std across 5 folds)

True \ Pred	Badger	Bird	Boar	Butterfly	Cat	Dog	Fox	Lizard	Cow	Porcupine	Weasel	Wolf
Badger	11.6±2.1	0.0±0.0	0.2±0.4	0.2±0.4	0.0±0.0	0.2±0.4	0.0±0.0	0.0±0.0	0.2±0.4	1.6±1.0	0.0±0.0	1.0±1.5
Bird	0.0±0.0	34.6±1.6	0.8±0.7	0.0±0.0	1.0±1.1	0.4±0.8	0.8±1.2	0.4±0.5	0.8±0.7	0.4±0.8	0.0±0.0	0.8±0.7
Boar	0.0±0.0	1.2±1.0	22.8±1.5	0.0±0.0	0.6±0.5	0.0±0.0	2.0±0.6	0.0±0.0	1.4±0.5	0.2±0.4	0.0±0.0	0.0±0.0
Butterfly	0.2±0.4	0.2±0.4	0.0±0.0	13.2±0.7	0.0±0.0	0.0±0.0	0.0±0.0	1.4±0.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Cat	0.0±0.0	0.4±0.5	0.0±0.0	0.2±0.4	13.2±1.3	0.0±0.0	0.4±0.8	0.0±0.0	0.4±0.5	0.0±0.0	0.0±0.0	0.4±0.5
Dog	0.0±0.0	0.2±0.4	0.4±0.5	0.4±0.5	0.0±0.0	12.4±2.3	0.0±0.0	1.0±0.6	0.2±0.4	0.0±0.0	0.0±0.0	0.4±0.5
Fox	0.4±0.5	1.2±1.2	0.6±0.8	0.0±0.0	0.2±0.4	0.4±0.5	7.8±1.5	0.2±0.4	1.4±1.4	0.6±0.8	0.2±0.4	2.0±1.3
Lizard	0.2±0.4	0.2±0.4	0.4±0.5	2.4±1.4	0.0±0.0	0.2±0.4	0.6±0.8	10.2±2.5	0.2±0.4	0.2±0.4	0.0±0.0	0.4±0.5
Cow	0.2±0.4	1.4±1.0	0.8±0.7	0.0±0.0	0.8±1.0	0.0±0.0	0.2±0.4	0.0±0.0	55.6±1.9	0.2±0.4	0.4±0.5	1.2±0.7
Porcupine	0.6±0.8	0.2±0.4	0.2±0.4	0.0±0.0	0.2±0.4	0.0±0.0	0.2±0.4	0.0±0.0	0.0±0.0	13.2±1.2	0.0±0.0	0.4±0.8
Weasel	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.4±0.5	0.0±0.0	0.0±0.0	0.0±0.0	0.2±0.4	0.0±0.0	14.4±0.8	0.0±0.0
Wolf	0.6±0.8	0.2±0.4	0.4±0.5	0.2±0.4	0.2±0.4	0.8±0.7	1.0±1.3	0.4±0.5	0.4±0.8	0.6±0.8	0.0±0.0	10.2±1.5

Table 3: Confusion matrix of Decision Tree (mean ± std over 5 folds)

True \ Pred	Badger	Bird	Boar	Butterfly	Cat	Dog	Fox	Lizard	Cow	Porcupine	Weasel	Wolf
Badger	14.8±0.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.2±0.4	0.0±0.0	0.0±0.0
Bird	0.0±0.0	39.2±1.2	0.0±0.0	0.2±0.4	0.0±0.0	0.0±0.0	0.2±0.4	0.4±0.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Boar	0.2±0.4	0.0±0.0	26.6±1.2	0.0±0.0	0.2±0.4	0.0±0.0	0.4±0.5	0.2±0.4	0.0±0.0	0.2±0.4	0.0±0.0	0.4±0.5
Butterfly	0.0±0.0	0.0±0.0	0.0±0.0	15.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Cat	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	15.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Dog	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	15.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Fox	1.0±0.9	0.2±0.4	0.2±0.4	0.2±0.4	0.2±0.4	0.0±0.0	12.2±2.1	0.0±0.0	0.0±0.0	0.2±0.4	0.0±0.0	0.8±1.2
Lizard	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	15.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Cow	0.0±0.0	0.0±0.0	1.0±1.3	0.0±0.0	0.0±0.0	0.0±0.0	0.4±0.5	0.2±0.4	58.6±1.4	0.0±0.0	0.0±0.0	0.6±0.8
Porcupine	0.2±0.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	14.8±0.4	0.0±0.0	0.0±0.0
Weasel	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	15.0±0.0	0.0±0.0
Wolf	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	15.0±0.0

Table 4: Confusion matrix of KNN ($k = 1$) (mean ± std over 5 folds)

True \ Pred	Badger	Bird	Boar	Butterfly	Cat	Dog	Fox	Lizard	Cow	Porcupine	Weasel	Wolf
Badger	14.2±1.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.2±0.4	0.0±0.0	0.6±1.2
Bird	0.0±0.0	39.2±1.2	0.0±0.0	0.2±0.4	0.0±0.0	0.2±0.4	0.2±0.4	0.2±0.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Boar	0.2±0.4	0.2±0.4	26.6±1.4	0.0±0.0	0.0±0.0	0.0±0.0	0.6±0.8	0.2±0.4	0.0±0.0	0.2±0.4	0.0±0.0	0.2±0.4
Butterfly	0.0±0.0	0.0±0.0	0.0±0.0	15.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Cat	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	14.8±0.4	0.0±0.0	0.0±0.0	0.2±0.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Dog	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	15.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Fox	0.8±1.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	12.6±1.6	0.2±0.4	0.0±0.0	0.4±0.5	0.0±0.0	1.0±1.1
Lizard	0.0±0.0	0.0±0.0	0.0±0.0	2.8±3.2	0.0±0.0	0.0±0.0	0.0±0.0	12.2±3.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Cow	0.0±0.0	0.0±0.0	0.2±0.4	0.0±0.0	0.0±0.0	0.0±0.0	0.2±0.4	0.0±0.0	59.8±1.2	0.4±0.5	0.0±0.0	0.2±0.4
Porcupine	1.2±0.7	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	13.2±1.5	0.0±0.0	0.6±0.8
Weasel	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	15.0±0.0	0.0±0.0
Wolf	0.6±1.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.2±0.4	0.0±0.0	0.6±0.8	0.6±0.8	0.0±0.0	13.0±1.7

Table 5: Confusion matrix of VitDet finetuned (frozen backbone) (mean ± std over 5 folds)

True \ Pred	Badger	Bird	Boar	Butterfly	Cat	Dog	Fox	Lizard	Cow	Porcupine	Weasel	Wolf
Badger	8.0±6.8	0.0±0.0	0.0±0.0	6.0±7.3	0.0±0.0	0.0±0.0	0.4±0.8	0.0±0.0	0.0±0.0	0.0±0.0	0.6±1.2	0.0±0.0
Bird	0.4±0.5	35.4±2.6	0.0±0.0	0.4±0.5	0.2±0.4	0.0±0.0	0.4±0.8	0.8±1.0	2.0±2.8	0.2±0.4	0.2±0.4	0.0±0.0
Boar	1.8±2.4	4.6±2.6	2.8±5.6	2.0±2.3	0.0±0.0	0.4±0.8	3.6±7.2	0.0±0.0	11.6±6.7	0.2±0.4	1.2±1.2	0.0±0.0
Butterfly	3.0±6.0	0.0±0.0	0.0±0.0	10.2±5.4	0.0±0.0	0.0±0.0	0.0±0.0	1.2±1.2	0.0±0.0	0.6±1.2	0.0±0.0	0.0±0.0
Cat	1.0±2.0	7.8±4.1	0.0±0.0	1.0±1.3	1.4±2.8	0.8±1.6	0.4±0.8	1.4±2.0	1.0±0.9	0.0±0.0	0.2±0.4	0.0±0.0
Dog	0.0±0.0	7.6±6.3	0.0±0.0	0.0±0.0	0.2±0.4	3.0±4.0	0.0±0.0	0.2±0.4	3.4±2.4	0.4±0.5	0.0±0.0	0.2±0.4
Fox	0.6±0.5	5.0±1.4	0.0±0.0	0.4±0.5	0.2±0.4	0.2±0.4	1.0±1.5	0.0±0.0	6.8±1.9	0.4±0.5	0.4±0.5	0.0±0.0
Lizard	1.6±2.7	5.2±2.3	0.0±0.0	3.2±2.8	0.2±0.4	0.4±0.8	0.2±0.4	2.8±3.9	0.2±0.4	1.0±1.3	0.2±0.4	0.0±0.0
Cow	0.2±0.4	10.0±6.8	1.0±2.0	0.4±0.5	0.2±0.4	0.2±0.4	1.6±3.2	0.0±0.0	47.2±4.3	0.0±0.0	0.0±0.0	0.0±0.0
Porcupine	3.2±3.3	4.0±3.9	0.2±0.4	3.0±3.7	0.0±0.0	0.4±0.8	0.8±1.6	0.0±0.0	1.0±0.9	2.2±1.9	0.2±0.4	0.0±0.0
Weasel	1.6±3.2	3.4±4.0	0.0±0.0	2.4±3.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.4±0.5	0.8±1.6	6.4±7.1	0.0±0.0
Wolf	2.4±2.3	2.8±2.6	0.4±0.5	1.6±2.3	0.0±0.0	0.6±0.8	0.2±0.4	0.8±1.2	4.0±1.5	0.2±0.4	0.6±0.8	1.4±2.8

Table 6: Confusion matrix of VitDet finetuned (mean ± std over 5 folds)