

Towards Zero-Shot Camera Trap Image Categorization

Jiří Vyskočil¹  and Lukas Picek^{1,2} 

¹ Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czechia

² Inria, LIRMM, Université de Montpellier, CNRS, Montpellier, France

Abstract. This paper describes the search for an alternative approach to the automatic categorization of camera trap images. First, we benchmark state-of-the-art classifiers using a single model for all images. Next, we evaluate methods combining MegaDetector with one or more classifiers and Segment Anything to assess their impact on reducing location-specific overfitting. Last, we propose and test two approaches using large language and foundational models, such as DINOv2, BioCLIP, BLIP, and ChatGPT, in a zero-shot scenario. Evaluation carried out on two publicly available datasets (WCT from New Zealand, CCT20 from the Southwestern US) and a private dataset (CEF from Central Europe) revealed that combining MegaDetector with two separate classifiers achieves the highest test accuracy. This approach reduced the relative error of a single BEiT2 classifier by approximately 42% on CCT20, 48% on CEF, and 75% on WCT. Besides, as the background is removed, the error in terms of accuracy in new locations is reduced to half. The proposed zero-shot pipeline based on DINOv2 and FAISS achieved competitive results (1.0% and 4.7% smaller on CCT20, and CEF, respectively), which highlights the potential of zero-shot approaches for camera trap image categorization.

Keywords: Camera Traps · Classification · Retrieval · BLIP · DINOv2 · Zero-shot · Vision and Language · ChatGPT · SAM · MegaDetector.

1 Introduction

Camera traps are valuable assets in ecological research. They are commonly used to estimate wildlife populations, species distribution, and their interactions [4, 24, 31, 38]. In many cases, the data (i.e., images) are still processed manually, which is extremely time-consuming, given the relatively high number of operated camera traps and their continuous data flow. Therefore, a concerted effort is being made to automate this process using machine learning and computer vision [22, 40]. While some studies have achieved human-level performance [37, 40], some challenges persist. For instance, models trained on a set of locations perform poorly in new ones, and models trained in the closed set setting must be re-trained. The same applies to new and unseen species, where adaptability is crucial, as it ensures that if a new, rare animal species appears in the monitored area, the classifier can recognize it.

In this paper, we test how existing foundational models perform in automatic camera trap image categorization and if they allow to overcome the above-mentioned problems. First, we evaluate state-of-the-art CNN- and Transformer-based classification architectures on three datasets (i.e., WCT [1] from New Zealand, CCT20 [3] from California, and CEF from Central Europe³) from different continents and with different species. Second, we use MegaDetector (MD) [21] and Segment Anything (SAM) [15] models for zero-shot detection and segmentation and test how these models improve classification performance and how they can help mitigate the issue of overfitting to the location/background. At last, we propose and test two approaches using large language and foundational models, such as DINOv2 [8,25], BioCLIP [35], BLIP [16], and ChatGPT [32], in a zero-shot scenario that perform similarly good in *seen* and *new locations* as the best approach based on supervised learning.

2 Related Work

In the past, camera trap image categorization was often done manually, which was neither effective nor fast [11, 36]. In response, ecologists and the machine learning community naturally shifted their interest in developing novel methods based on machine learning and computer vision.

Pioneer studies in automatic camera trap image categorization [12, 43] were based on local features (e.g., SIFT, SURF) or early adoptions of neural network classifiers. With the fast progress in machine learning, especially CNNs, many studies [20, 23, 40, 41] focused on fine-tuning standardized architectures for image categorization such as ResNet and VGG. Recent work [7, 13, 22, 33, 34] employs a two-step process involving object detection⁴ followed by classification. This approach primarily mitigates information loss caused by resizing images to fit the expected classifier input and consequentially reduces overfitting to training locations. Additionally, detecting the animals has clear advantages, as it enables handling multi-species presence and counting the animals present, in addition to categorization.

Even though there is a long track record of methods for camera trap image categorization, most existing approaches share the following drawbacks: (i) Methods that do not use object detection to crop the animal before classification tend to overfit to the background, resulting in poor performance on new and *unseen* locations. Additionally, even with the cropped animal, some background pixels are present and, therefore, can result in overfitting. (ii) All available pre-trained models, trained on a closed set of categories (i.e., species), cannot be effectively deployed in different climates or continents where species not present in the training data are naturally present. (iii) Existing datasets lack standardization, and available categories are usually on a different taxonomical level, i.e., species, genus, family.

³ Due to animal safety concerns this dataset is available only after signing a Non-Disclosure Agreement (NDA).

⁴ Usually using large pre-trained models, e.g., MegaDetector [21].

3 Datasets

While selecting the datasets for our experiments, we considered the different geographical locations and species. Therefore, we evaluated our experiments on three datasets that originate from North America (Caltech Camera Traps-20), Oceania (Wellington Camera Traps), and Europe (Central Europe Fauna-22).

We use CCT20 as the primary subject in our ablations due to its smaller scale (i.e., faster training) and wide usage. To verify the outcomes in different settings and geographic locations, we use the WCT and CEF22 datasets. Below, we briefly describe each dataset, with statistics provided in Table 1.

Note: *Typically, there is only one animal per image, but in rare cases where multiple species appear, we remove those images from the evaluation. Besides and similarly as in the original work [3], we also remove images without animals.*

Table 1: Statistics for selected camera trapping datasets. [†]Denotes custom split.

| Dataset | Boxes | Categories | Training | Validation | Test |
|--|-------|------------|----------|------------|--------|
| Caltech Camera Traps [3] | ✓ | 16 | 13,553 | 5,209 | 39,102 |
| (our) Central Europe Fauna | – | 32 | 92,603 | 26,479 | – |
| Wellington Camera Traps [†] [1] | – | 17 | 140,035 | 35,009 | 95,406 |

The **Caltech Camera Traps-20 (CCT20)** [3] dataset consists of images captured from 20 locations in the Southwestern United States. This dataset includes approximately 58,000 images, covering 15 animal categories as well as one "empty" image category. The data are neither balanced nor filtered, reflecting a more natural distribution of species across both time and location. To allow evaluation of location/environment regularization, the validation set is divided into two subsets: *cis* and *trans*. The *cis* subset includes only locations included in the training set, while the *trans* subset includes just new locations.

The **Wellington Camera Traps (WCT)** [1] dataset originates from 187 locations in New Zealand and contains around 270,000 images across 16 categories, plus one category for empty and unclassifiable images. Since the dataset is not pre-split into train, validation, and test subsets, we allocated 1/3 of the locations for testing and the remaining 2/3 for development, which includes both training and validation. Within the development set, the images are randomly divided in an 80/20 ratio for training and validation, respectively. This approach ensures a balanced distribution for model development and evaluation.

Unlike the previous two datasets, the **Central Europe Fauna-22 (CEF22)** is a private dataset originating from three different trapping projects and multiple distinct regions within Central Europe. The dataset consists of approximately 120,000 images and includes 32 species categories⁵. Notably, all empty images were prefiltered and are not included in the dataset. This curated approach focuses exclusively on images containing identifiable species, enhancing the dataset's utility for training and evaluation purposes.

⁵ Except two taxon labels (e.g., Aves and Martes) data are labeled with species labels.

4 Baseline Performance

To the best of our knowledge, there is no standardized benchmark for camera trap categorization (including the three selected datasets); therefore, we provide a few baseline experiments that evaluate state-of-the-art CNN- and Transformer-based architectures for image classification. This section presents multiple experiments and ablations that focus on benchmarking performance for CNN- and transformer-based classifiers. In the following section, these models are further enhanced with existing foundational models in a zero-shot detection and segmentation context. In Figure 1, we illustrate four designed approaches.

Evaluation protocol: In all the experiments and ablations presented in this section, we use the Top1 and Top3 accuracy and macro averaged F1 score ($F1_m$) as the evaluation metrics. The evaluation is primarily carried out on CCT20 validation subsets (i.e., `cis_val` + `trans_val`) and further tested on CEF22 and WCT datasets to verify the results transferability.

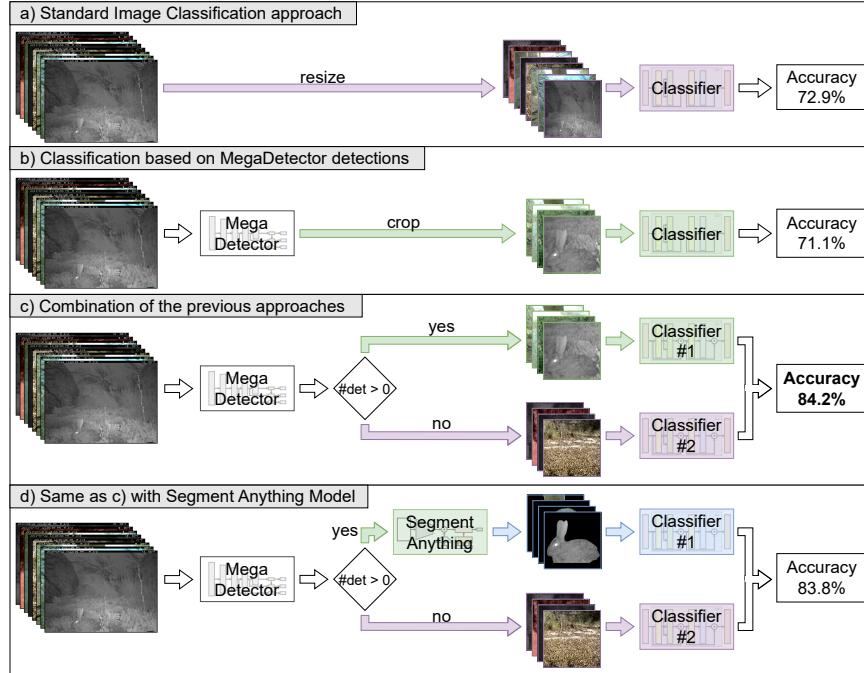


Fig. 1: Four baseline approaches illustration. (a) Standard image classification approach where images are resized and fed to a trained image classifier. (b) A more recent two-stage approach with object detection before the classification. This approach can suffer from missing reject options for images with no detection. (c) With two image classifiers, the problem of (b) can be easily mitigated. (d) Excluding background pixels using SAM could help to prevent overfitting to the location.

Benchmarking image classifiers: Based on the results reported in general image classification [5, 27] and species classification [28, 29], we primarily focus on transformers, but we add ConvNext and ResNext into the mix to compare it for camera trap image analysis.

Experiment settings: We use pre-trained models from the timm python library [39], version 0.9.12. We train all models using SGD optimizer and the momentum of 0.9 [30] for 40 epochs with a batch size of 64, and the cosine scheduler, which decreases the learning rate from $1e^{-3}$ to $2.54e^{-6}$, and the following augmentations during the training: *RandomResizeCrop* with a scale from 0.8 to 1.0, and RandAugment [6] with a magnitude of 0.2.⁶

Results: Achieved results are reported in Table 2 and Table 3. As expected, transformers and CNNs performed competitively, but transformers achieved better performance in new locations. All models underperformed significantly on the unseen *trans locations*, while the CNNs exhibit a higher score difference between the *trans* and *cis*, indicating that they overfit more on the locations themselves. For the following experiments, we will use just the top 3 architectures, i.e., BEiT, BEiT2, and EfficientViT.

Table 2: Ablation on model architecture. We benchmark selected CNN- and transformer-based architectures on the CCT20 val. sets without empty images. The ranking in camera trap image categorization more-less follows the ranking on ImageNet.

| Architecture | Pre-train checkpoint | Input size | <i>Cis location</i> | | <i>Trans location</i> | | <i>Both</i> |
|----------------------|-------------------------|------------------|---------------------|-----------------|-----------------------|-----------------|-------------|
| | | | Top1 | F1 _m | Top1 | F1 _m | Top1 |
| ConvNeXt-Base [19] | IN22k | 224 ² | 84.9 | 68.2 | 36.2 | 20.5 | 65.7 |
| ResNeXt-50 [42] | IN1k | 224 ² | 86.8 | 69.7 | 35.8 | 16.4 | 66.7 |
| EfficientViT-B3 [5] | IN1k | 224 ² | 78.9 | 61.2 | 40.9 | 19.4 | 63.9 |
| ViT-Base/p16 [9] | IN1k | 224 ² | 84.6 | 68.0 | 49.4 | 20.9 | 70.7 |
| SwinV2-Base/w16 [17] | IN1k | 256 ² | 85.4 | 66.9 | 49.1 | 21.9 | 71.1 |
| Swin-Base/p4-w7 [18] | IN22k | 224 ² | 83.6 | 67.0 | 52.0 | <u>24.3</u> | 71.1 |
| BEiT-Base/p16 [2] | IN22k | 224 ² | 84.4 | 67.2 | <u>54.4</u> | 21.1 | 72.6 |
| BEiT2-Base/p16 [27] | IN22k | 224 ² | <u>85.9</u> | <u>68.6</u> | 53.0 | 21.7 | <u>72.9</u> |
| EfficientViT-L3 [5] | IN1k | 224 ² | 83.5 | 67.0 | 60.4 | 28.6 | 74.4 |

Table 3: Transformers performance. All models achieved a competitive accuracy on all three datasets. However, the BEiT2 model performed much better on the tail species. *fps was measured on NVIDIA A40 (GPU) and AMD EPYC 7543 (CPU).

| Architecture | CCT20 dataset | | | CEF22 dataset | | | WCT dataset | | | fps* |
|-----------------|---------------|-------------|-----------------|---------------|-------------|-----------------|-------------|-------------|-----------------|------------|
| | Top1 | Top3 | F1 _m | Top1 | Top3 | F1 _m | Top1 | Top3 | F1 _m | |
| BEiT-Base/p16 | 72.6 | <u>89.7</u> | 60.1 | 85.3 | 94.5 | <u>68.4</u> | 86.5 | 98.0 | <u>72.7</u> | 331 |
| BEiT2-Base/p16 | <u>72.9</u> | 89.6 | <u>61.9</u> | 87.5 | 95.5 | 72.2 | <u>86.0</u> | 98.8 | 75.5 | 333 |
| EfficientViT-L3 | 74.4 | 90.8 | 62.7 | <u>85.7</u> | <u>94.9</u> | 62.7 | 85.1 | <u>98.7</u> | 67.1 | 295 |

⁶ Selected as sub-optimal for all selected models based on our preliminary experiments.

5 Zero-shot Detection and Segmentation

This section analyzes and provides a qualitative and quantitative evaluation of the potential of two pre-trained foundational models, the MegaDetector (MD) [21] and the Segment Anything Model (SAM) [15], in processing camera trap images. Furthermore, we explore decision-making strategies when MD detects no object, including a) considering an image to be empty and b) classifying the whole image. At last, leveraging the insights gained from our ablations, we test three state-of-the-art transformer-based architectures in combination with MD and SAM trained on three distinct datasets with different species and origins.

Image data processing: We run MD to obtain detections of objects for all images. Then crop the detections from the image so that the resulting image has the same width and height to avoid breaking the aspect ratio. Furthermore, we use the SAM (initialized with the MD’s detection) to remove background pixels and move the object to the center. See Figure 2 for examples of the detected and segmented objects in the CCT20 dataset.

Classification models: Since neither MD nor SAM provides class categories, we will use the top three architectures (in terms of accuracy) from the previous ablation. We have four different types of classifiers, each differing in the expected input: (a) full-size images, (b) cropped detections, (c) cropped detections with segmented objects, and (d) both full-size and cropped images. We investigate combinations of these in our experiments and ablation studies.

Building a classifier for inference: Once the models are trained, we create a predictor containing two classifiers. First of all, we apply MD on the image. If MD finds an object, we use classifier #1, which is trained to recognize objects from cropped or even segmented images. If no object is found, we use classifier #2 on the whole image. Besides, we perform an experiment in which we use only one classifier on cropped and full-size images.

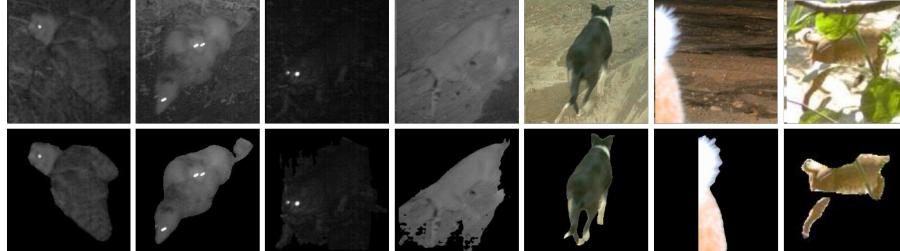


Fig. 2: Zero-shot segmentation with Segment Anything. Random samples from CCT20 datasets are processed by MegaDetector, and resulting detections are fed into SAM. Even with the poor quality of the data, the zero-shot segmentation performs relatively well across a wide range of species. However, with the infra-red images and the small size of an object, the SAM starts to fail (3rd and 4th column from the left).

5.1 Performance Evaluation

We compare three types of classifiers that differ in their expected input data to explore the benefits of using MD and/or SAM on camera trap images. Each classifier expects one of the following inputs: (a) resized images, (b) images cropped based on detection from MegaDetector, and (c) cropped detections with removed background (we use Segment Anything Model, which is fed with the MegaDetector detections). The comparison is made on a filtered validation set of the CCT20 dataset that contains only images with objects. By cropping detections from images, we have improved the Top1 accuracy on average by 10.3%, 5.2%, and 10% on the CCT20, CEF22, and WCT datasets, respectively. The improvement is lower on the dataset with European fauna as it was most likely not used in the training of the MD. Using the segmented animals, the performance decreased by an average of 5.1% on CCT20, 2.8% on CEF22, and 0.8% on WCT. See Table 4 for more comprehensive evaluation.

Location overfitting: We use the CCT20 *cis* and *trans* subsets to test the hypothesis that image classifiers tend to overfit to background pixels and, therefore, perform poorly in new locations. Following the results listed in Table 4, it is evident that reducing the number of background pixels available for training with both cropping our segmenting the object reduces the location overfitting. Besides training the model on cropped objects, increases the overall classification performance as more detail is provided. Using MD and "feeding" a classifier with cropped detections still results in around 15% lower performance in the new locations. However, it is a significant improvement as without the cropping, the resulting performance in *trans* (*unseen*) locations dropped by roughly 30% in terms of Top1 accuracy, and the error almost doubled.

Table 4: Ablations on location overfitting. We compare the performance in terms of Top1 Accuracy of three models trained on (a) resized images, (b) cropped animals, and (c) cropped + segmented animals on three datasets. While (a) shows huge overfitting to the location, the (b) and (c) approaches show much better generalization and robustness to changes in testing location. In general, (b) achieves the best performance.

| Architecture | MD | SAM | | | | CCT | |
|-----------------|----|-----|-------------|-------------|-------------|-------------|--------------|
| | | | CCT | CEF | WCT | <i>cis</i> | <i>trans</i> |
| BEiT-Base/p16 | – | – | 72.6 | 85.3 | 86.5 | 84.4 | 54.4 |
| BEiT-Base/p16 | ✓ | – | 83.9 | 91.1 | 96.0 | 89.6 | 75.2 |
| BEiT-Base/p16 | ✓ | ✓ | 80.8 | 88.8 | 95.0 | 87.3 | 70.8 |
| BEiT-Base/p16 | – | – | 72.9 | 87.5 | 86.0 | 85.9 | 53.0 |
| BEiT-Base/p16 | ✓ | – | 84.2 | 92.2 | 96.6 | 90.1 | 75.0 |
| BEiT-Base/p16 | ✓ | ✓ | 83.8 | 89.2 | 95.8 | 89.2 | 75.5 |
| EfficientViT-L3 | – | – | 74.4 | 85.7 | 85.1 | 83.5 | 60.4 |
| EfficientViT-L3 | ✓ | – | 83.1 | 90.8 | 95.0 | 87.2 | 76.9 |
| EfficientViT-L3 | ✓ | ✓ | 81.1 | 87.6 | 94.4 | 85.5 | 74.2 |

5.2 Additional Ablations

Dealing with images without detections: Although MD is a powerful foundational model, it has its limitations, such as generating false positives or failing to detect objects altogether. Therefore, we provide an additional ablation that compares two classification methodologies when MD does not detect any objects: (a) using a single classifier for both cropped and full-size images and (b) using separate classifiers for each.

Using a single classifier (a) over the two separate classifiers (b) reduces computational complexity, but it reduces the Top1 accuracy by around 1% for all three architectures. For example, EfficientViT improves accuracy by 1.4% with two classifiers, as shown in the Table 5.

What to do with "empty" images? A common issue in camera trapping is false sensor activation, leading to saving "empty" images with no animals. In this ablation, we test two approaches to distinguish empty images from those with animals. We use the CCT20 dataset, which includes an "empty" category in the validation and test sets, though not in the training set.

We test the following two approaches:

- (a) Treat images as "empty" when MD detects no object.
- (b) Generate or add object-free data to the training set and train a second classifier on full-size images. We generate average images based on location, date, and time.

Results in Table 6 show that assuming an image is empty when MD does not detect anything significantly reduces performance by up to 18.5%, highlighting the need for better strategies in handling empty images.

Table 5: Classification of images with none MegaDetector detections. Performance of (a) one classifier for cropped images from detections and full-size images and (b) two separated classifiers, one trained on cropped images and the second one on full-size images.

| Architecture | Classifier | Top1 |
|------------------|------------|-------------|
| BEiT-Base/p16 | (a) | 66.5 |
| BEiT-Base/p16 | (b) | <u>67.5</u> |
| BEiT-V2-Base/p16 | (a) | 67.4 |
| BEiT-V2-Base/p16 | (b) | 68.2 |
| EfficientViT-L3 | (a) | 65.8 |
| EfficientViT-L3 | (b) | 67.2 |

Table 6: Dealing with empty images. We compare a method that trusts the MD to declare the image empty if no object is found with another that uses two classifiers: 1st on a cropped image if an object is detected and 2nd on the full image if not.

| If no detection, then | Top1 |
|--|---------------------|
| consider it as <i>empty image</i> . use 2 nd classifier. | 49.1 <u>67.5</u> |
| consider it as <i>empty image</i> . use 2 nd classifier. | 49.7 68.2 |
| consider it as <i>empty image</i> . use 2 nd classifier. | 50.1 67.2 |

6 Zero-shot Classification

In addition to the standard classification approach with trained closed-set model for camera trap image categorization, and following the success of LLMs, we further explore the capabilities of existing foundational models for multi-modal processing (e.g., BLIP [16] and ChatGPT [32]) and image retrieval using Bio-CLIP [35] and DINOv2 [25]. For "classification" based on image retrieval, we use the FAISS library [10, 14] developed for efficient similarity search and clustering of deep embeddings. Both evaluated pipelines are illustrated in Figure 3.

Note: We are aware of the limited reproducibility when using ChatGPT, but one of the goals of this work is to fully explore what are the current LLMs capabilities in scenarios related to ecology compared to standardly used methods.

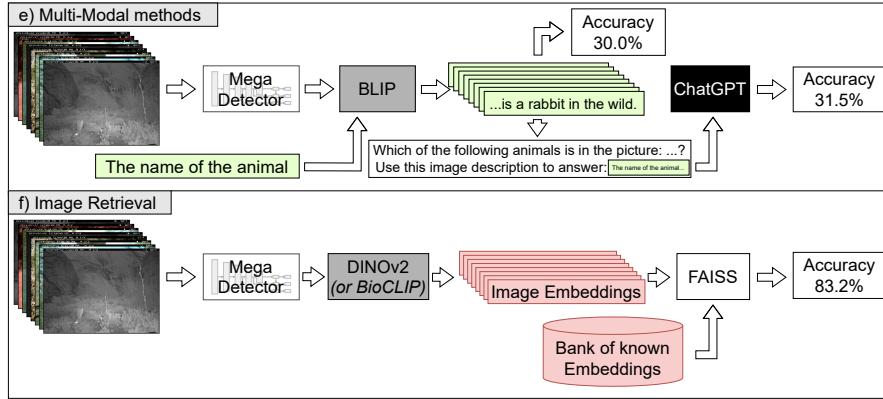


Fig. 3: Zero-Shot approaches. (e) Multi-modal methods extract image info. based on given textual and image prompts. (f) Embeddings are generated from the training set, and during inference, selective search finds similar images from the database.

6.1 Large Language and Multi-Modal Models

We use an image captioning model, BLIP, that provides a textual output to a prompt consisting of text and image. To enhance BLIP's focus on animals, we condition various image captions by providing text that includes keywords such as "what," "species," "animal," etc. Furthermore, we use ChatGPT (based on GPT-4) as a filter to identify animals from the given options and image description. We consider the image to be empty if none or more than one of the possible categories is included in the response.

BLIP [16]: We test how the pre-trained multi-modal image captioning model's (i.e., BLIP) performance depends on different conditional captioning inputs (e.g., "*The species of the animal is...*", or "*The name of the animal...*"). Overall the performance fluctuated inconsistently across the datasets and "prompts," and

conditioning the output with no captions can yield relatively good scores. Using "*A running...*" as captioning yielded the best results on the CEF22 datasets; however, slightly underperformed on the CCT20 and WCT. Still, this captioning seems to be a cross-dataset sub-optimum. Interestingly, when the model fails to recognize an animal, it often defaults to "a bear in the wild," even if no bear is present in the dataset, or to phrases like "is not visible"/"is on the camera screen." For detailed performance, see Table 7. Compared to baseline performance with approach (a) and Top1 accuracy of BEiT2-Base/p16 (i.e., 84.2, 92.2, and 96.6 on CCT20, CEF22, and WCT respectively), BLIP underperformed badly, achieving roughly half of the accuracy.

Table 7: Ablation on "captioning prompting" of BLIP.

| Conditional image captioning | CCT20 | CEF22 | WCT |
|---|-------------|-------------|-------------|
| – | 21.7 | 32.6 | <u>30.1</u> |
| <i>The picture shows a cute ...</i> | 14.7 | 35.6 | 21.8 |
| <i>I see cute ...</i> | 19.0 | 36.6 | 30.2 |
| <i>The species of the animal is ...</i> | 20.0 | 2.5 | 18.0 |
| <i>The animal in the picture is ...</i> | 20.1 | 1.9 | 16.8 |
| <i>A running ...</i> | 21.5 | 40.9 | 27.8 |
| <i>A peeking ...</i> | 22.5 | <u>38.1</u> | 29.2 |
| <i>This animal is called ...</i> | <u>24.4</u> | 28.3 | 24.9 |
| <i>The name of the animal ...</i> | 24.9 | 37.3 | 25.7 |

Fine-tuning BLIP captions with ChatGPT: ChatGPT has gained popularity due to its suitability across diverse topics and its ability to engage in human-like conversations. However, its capabilities in camera trap image categorization have not been explored yet. Since the ChatGPT API only allows textual inputs, we use it to further process the BLIP's outputs given the image and sub-optimal conditional captioning and put it as an input into the GPT-4 based model [32]. The proposed textual input for the ChatGPT is as follows:

Write a one-word answer to this question: "Which of the following animals is in the picture: <list of categories>?" Consider this image description in the answer: <BLIP generated caption>.

Where *<list of categories>* are all targets without "empty" category, separated with commas and *<BLIP generated caption>* is BLIP's output.

Based on our experiment, ChatGPT can slightly improve the performance of BLIP, with an improvement of approximately 1.5% on the CCT20 dataset. Our observations indicate that, in some cases, ChatGPT adjusts the final prediction by selecting a class from the options provided. However, more frequently, ChatGPT ignores the options and instead propagates the species identified in the caption generated by BLIP, even if it is not among the provided options (e.g., "bear"). Furthermore, ChatGPT tends to list the options from which it claims to have selected its answer, even when the chosen answer is absent from the list.

6.2 Retrieval-based Classification

In this section, we test the suitability of existing pre-trained foundational models (e.g., BioCLIP [35] and DINOv2 [8, 25]) for zero-shot camera trap image categorization in a retrieval-like setting. Zero-shot classification of camera trap images can solve some of the major drawbacks of the existing approaches as it allows the classification of images into categories not seen during training by retrieving and comparing them against a set of predefined class examples. This means that the database can be extended by new species at any time without re-training the model. Furthermore, this approach allows simple updates of the class reference samples, so weekly labeled data and data with no additional value might be easily filtered out or replaced. On the other hand, after the inference, one more step, i.e., similarity search, is needed. For example predictions, see Figure 4.

Method: First, we feed forward all the training images into the image encoders and store image embeddings into the *database*. Second, we generate embeddings for all *query* data the same way. At last, we use FAISS library [10, 14] to perform an efficient similarity search. In both cases, the images are resized to 256×256 and center-cropped to 224×224 . As a baseline, we use the standard L2 distance without feature vector normalization and k-NN classifier with $k = 1$ to select the species prediction.

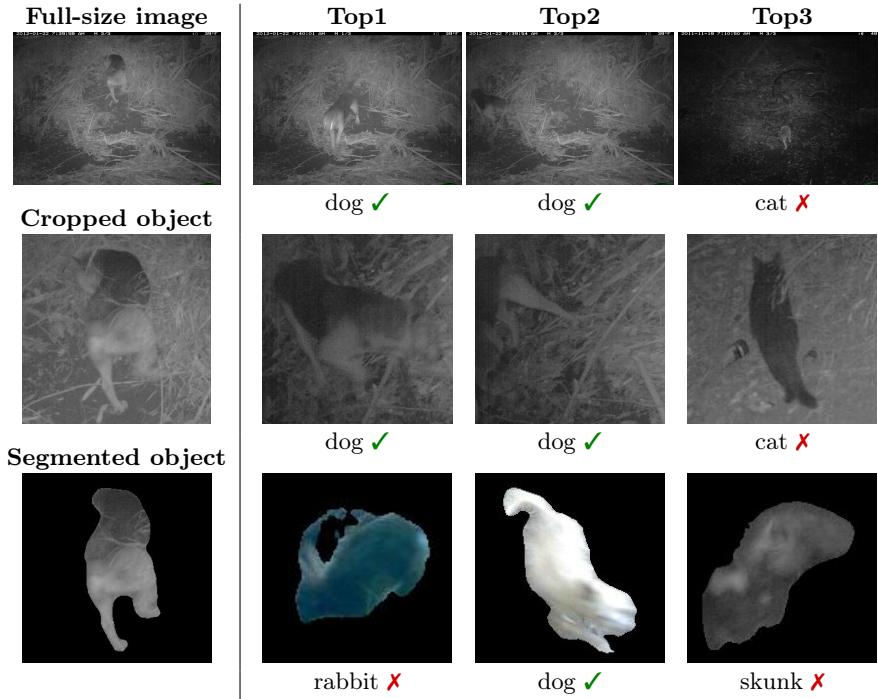


Fig. 4: Top3 closest images to given inputs using DINOv2_G embeddings.

Results: Overall, the best-performing model was the DINOv2_G, which achieved an outstanding performance comparable to models trained with full supervision. More precisely, the DINOv2_G achieved a Top1 accuracy of 83.2% and 87.5% on CCT and CEF datasets, respectively. Interestingly, the newer and "*supposed-to-be better*" DINOv2_B+reg [8] slightly underperformed the vanilla DINOv2_B and the BioCLIP_B heavily underperformed both DINO models.

As in previous experiments, cropping detected animals was also beneficial for image retrieval, and the segmentation seems to be unnecessary. In the scenario with just a single classifier, the DINOv2_G performed similarly on the CEF dataset and slightly better on the CCT dataset. A more comprehensive qualitative and quantitative evaluation is listed in Table 8, and Figure 4.

Table 8: Camera trap image categorization in retrieval-like setting. We compare BioCLIP, DINOv2 [25], and DINOv2+reg [8], on the CCT20 and CEF20 datasets in terms of Top1 accuracy. To allow direct comparison with the *standard* approaches (i.e., (a), (b), and (c) in Figure 1), we add scores achieved with BEiT_{v2} classifiers explicitly trained to classify camera trap images on those datasets.

| | MD | SAM | BioCLIP _B | DINOv2 _B +reg | DINOv2 _B | DINOv2 _G | BEiT _{v2B} |
|-----|----|-----|----------------------|--------------------------|---------------------|---------------------|---------------------|
| CCT | – | – | 60.5 | 65.5 | 67.8 | 74.4 | 72.9 |
| | ✓ | – | 65.4 | 77.5 | 77.8 | 83.2 | 84.2 |
| | ✓ | ✓ | 62.7 | 71.8 | 70.2 | 75.6 | 83.8 |
| CEF | – | – | 70.1 | 80.9 | 81.5 | 87.1 | 87.5 |
| | ✓ | – | 69.9 | 82.5 | 82.6 | 87.5 | 92.2 |
| | ✓ | ✓ | 61.6 | 74.2 | 75.3 | 81.1 | 89.2 |

Ablation on location generalization. Given that image retrieval-based classifiers demonstrated comparable results to fine-tuned models, we performed an additional ablation on three African datasets⁷ from Snapshot Safari [26]: Enonkishu (ENO), Kgalagadi (KGA), and Kruger (KRU). As these datasets are not split into development and testing subsets, we use images from the first x locations⁸ to build the database and the remaining locations for testing. In other words, the results reported in Table 9 for ENO, KGA, and KRU are based solely on images from unseen locations.

Retrieval-based methods achieve results comparable to, and in some cases exceeding, the BEiT_{v2} fine-tuned models. This is particularly evident on the African datasets, which contain relatively limited training data for classification (3,929, 662, and 1,312 images for ENO, KGA, and KRU, respectively), where model training completely failed for KRU. Considering that the performance of retrieval-based methods can be easily improved by simply adding new data

⁷ We remove empty images and compare models without using MD or SAM.

⁸ We use the first 8, 15, and 26 locations for the ENO, KGA, and KRU datasets, resulting in 20%, 31%, and 29% of the total images being used for testing, respectively.

Table 9: Results of image retrieval-based classifier on additional datasets.

| Model | CCT | CEF | WCT | ENO | KGA | KRU |
|----------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| BioCLIP _B | 60.5 | 70.1 | 80.3 | 44.3 | 72.8 | 55.7 |
| DINOv2 _G | 74.4 | <u>87.1</u> | <u>83.6</u> | 65.9 | <u>79.9</u> | 79.6 |
| BEiT _{V2B} | <u>72.9</u> | 87.5 | 86.0 | <u>53.5</u> | 83.0 | 41.1 |

to the database, whereas standard image classifiers require model fine-tuning, retrieval-based approaches prove promising results, especially for small datasets.

Ablation on matching strategy. In this experiment, we compare two similarity measures (L_2 norm and cosine similarity) and various k values for the k-NN classifier over DINOv2_G features on five distinct datasets. The results are inconclusive as the best-performing approach is different for each dataset. Therefore, we recommend testing the best approach before using it on your dataset. For further details, refer to Table 10.

Table 10: Matching strategy settings ablation. We compare how different settings affect DINOv2_G retrieval performance using L_2 norm and cosine similarity based on class centroids and various k values in the k-NN algorithm.

| Dataset | Cosine similarity | | L_2 norm | | | | |
|---------|-------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | 1-NN | 3-NN | 1-NN | 3-NN | 5-NN | 10-NN | centr. |
| CCT | 74.4 | 72.7 | 74.4 | <u>74.1</u> | 72.3 | 71.8 | 54.8 |
| CEF | 87.0 | <u>87.7</u> | 87.1 | 87.8 | 87.8 | <u>87.7</u> | 74.2 |
| ENO | 66.6 | 68.6 | 65.9 | 67.9 | <u>68.4</u> | 65.9 | 54.6 |
| KGA | 80.6 | 78.5 | 79.9 | 79.5 | <u>79.2</u> | <u>83.4</u> | 86.9 |
| KRU | <u>80.0</u> | 79.8 | 79.6 | 76.1 | 77.2 | 80.7 | 74.6 |

7 Conclusion and Perspectives

This paper presents a comprehensive comparative analysis of various approaches to automating camera trap image categorization. Our findings highlight the usefulness of integrating two independent classifiers—one specialized for cropped animals and the other for full images—with MegaDetector. This approach, where the appropriate classifier is selected based on whether the MegaDetector detects an animal, resulted in great improvements and allowed very limited overfitting to the location. More precisely, the combination of the MegaDetector with two fine-tuned BEiT_{V2} models reduced the relative error by around 42%, 48%, and 75% on the CCT20, CEF22, and WCT test sets, respectively.

In addition, we proposed two alternative approaches for zero-shot classification based on (i) multi-modal models (e.g., BLIP and ChatGPT) and (ii) image retrieval based on deep features from DINOv2 and BioCLIP. For both, we used MegaDetector to detect and crop the animals.

Interestingly, DINOv2_G achieved Top1 accuracy of 83.2% and 87.5% on CCT and CEF, respectively, trailing behind the best approach by 1.0% and 4.7% percentage points. This approach, leveraging robust visual representations and efficient similarity search, offers interesting properties such as (i) no requirement for fine-tuning while introducing new species and (ii) robustness toward location.

The provided analysis underscores the value of using specialized classifiers tailored to different aspects of the image. When initial object detection fails, integrating a secondary classifier for full images substantially enhances the classification of images that MegaDetector deems empty, thus reducing overall classification errors. Besides, our results achieved in a zero-shot scenario showed a promising direction for zero-shot camera trap image classification methods. Future work should focus on developing those methods, as they allow for straightforward adaptability for new species and locations.

The main outcomes derived from this work include the following:

MegaDetector rules. The high accuracy and robustness in detecting animals, people, and vehicles in camera trap images make MegaDetector an excellent asset for preprocessing raw camera trapping data. Since CNN- and Transformer-based classifiers require resizing the original images to an expected input size, the detection and cropping of animals from the original images reduce the data loss related to resizing and subsequently results in a considerable increase in performance and also a better regularization towards location changes.

Large language and multi-modal models do not perform well. While powerful in natural language processing and image captioning, ChatGPT and BLIP perform poorly on camera trap images. This is most likely due to their utility being confined to text- or text+image-based applications, making them unsuitable for the visual analysis of camera trap data.

A great potential of retrieval-like classification. DINOv2 features used in image retrieval settings for the classification of camera trap data are surprisingly robust, especially when using larger architectures. Comparing it with traditional classifier-based methods offers a promising alternative with similar performance and multiple benefits, e.g., new species or locations can be quickly and easily introduced just by adding a few samples into the *database*, and no fine-tuning is ever needed, which saves a considerable amount of CO₂.

Acknowledgements

This research was supported by the Technology Agency of the Czech Republic, project No. SS05010008, and by the grant of the University of West Bohemia, project No. SGS-2022-017. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. We also thank Friends of the Earth, Czechia, and Národní Park Šumava for providing the data to form the CEF dataset.

References

1. Anton, V., Hartley, S., Geldenhuis, A., Wittmer, H.U.: Monitoring the mammalian fauna of urban areas using remote cameras and citizen science. *Journal of Urban Ecology* **4**(1), juy002 (2018)
2. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
3. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European conference on computer vision (ECCV). pp. 456–473 (2018)
4. Blount, J.D., Chynoweth, M.W., Green, A.M., Şekercioğlu, Ç.H.: Covid-19 highlights the importance of camera traps for wildlife conservation research and management. *Biological Conservation* **256**, 108984 (2021)
5. Cai, H., Gan, C., Han, S.: Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. arXiv preprint arXiv:2205.14756 (2022)
6. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
7. Cunha, F., dos Santos, E.M., Colonna, J.G.: Bag of tricks for long-tail visual recognition of animal species in camera-trap images. *Ecological Informatics* **76**, 102060 (2023)
8. Dariseti, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvassy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. arXiv preprint arXiv:2401.08281 (2024)
11. Fegraus, E.H., Lin, K., Ahumada, J.A., Baru, C., Chandra, S., Youn, C.: Data acquisition and management software for camera trap data: A case study from the team network. *Ecological Informatics* **6**(6), 345–353 (2011)
12. Harris, G., Thompson, R., Childs, J.L., Sanderson, J.G.: Automatic storage and analysis of camera trap data. *Bulletin of the Ecological Society of America* **91**(3), 352–360 (2010)
13. Henrich, M., Burgueño, M., Hoyer, J., Haucke, T., Steinhage, V., Kühl, H.S., Heurich, M.: A semi-automated camera trap distance sampling approach for population density estimation. *Remote Sensing in Ecology and Conservation* **10**(2), 156–171 (2024)
14. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. *IEEE Transactions on Big Data* **7**(3), 535–547 (2019)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
16. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
17. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings

- of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
 19. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
 20. Miao, Z., Gaynor, K.M., Wang, J., Liu, Z., Muellerklein, O., Norouzzadeh, M.S., McInturff, A., Bowie, R.C., Nathan, R., Yu, S.X., et al.: Insights and approaches using deep learning to classify wildlife. *Scientific reports* **9**(1), 8137 (2019)
 21. Microsoft: Pytorch-wildlife: A collaborative deep learning framework for conservation. <https://github.com/microsoft/CameraTraps/blob/main/megadetector.md> (2023)
 22. Norouzzadeh, M.S., Morris, D., Beery, S., Joshi, N., Jojic, N., Clune, J.: A deep active learning system for species identification and counting in camera trap images. *Methods in ecology and evolution* **12**(1), 150–161 (2021)
 23. Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J.: Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences* **115**(25), E5716–E5725 (2018). <https://doi.org/10.1073/pnas.1719367115>, <https://www.pnas.org/content/115/25/E5716>
 24. O’Connell, A.F., Nichols, J.D., Karanth, K.U.: Camera traps in animal ecology: methods and analyses, vol. 271. Springer (2011)
 25. Oquab, M., Dariseti, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)
 26. Pardo, L.E., Bombaci, S., Huebner, S.E., Somers, M.J., Fritz, H., Downs, C., Guthmann, A., Hetem, R.S., Keith, M., Roux, A.I., et al.: Snapshot safari: A large-scale collaborative to monitor africa’s remarkable biodiversity. *South African Journal of Science* **117**(1-2), 1–4 (2021)
 27. Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F.: Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366* (2022)
 28. Picek, L., Šulc, M., Matas, J., Jeppesen, T.S., Heilmann-Clausen, J., Laessøe, T., Frøslev, T.: Danish fungi 2020-not just another image recognition dataset. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1525–1535 (2022)
 29. Picek, L., Šulc, M., Patel, Y., Matas, J.: Plant recognition by ai: Deep neural nets, transformers, and knn in deep embeddings. *Frontiers in plant science* **13**, 787527 (2022)
 30. Qian, N.: On the momentum term in gradient descent learning algorithms. *Neural networks* **12**(1), 145–151 (1999)
 31. Rowcliffe, J.M., Field, J., Turvey, S.T., Carbone, C.: Estimating animal density using camera traps without the need for individual recognition. *Journal of Applied Ecology* pp. 1228–1236 (2008)
 32. Ruu3f: Freegpt (2023), <https://github.com/Ruu3f/freeGPT>
 33. Schneider, D., Bellafkir, K.L.M.V.H., Farwig, M.M.N., Freisleben, B.: Recognizing european mammals and birds in camera trap images using convolutional neural networks (2023)

34. Shepley, A., Falzon, G., Meek, P., Kwan, P.: Automated location invariant animal detection in camera trap images using publicly available data sources. *Ecology and Evolution* **11**(9), 4494–4506 (2021)
35. Stevens, S., Wu, J., Thompson, M.J., Campolongo, E.G., Song, C.H., Carolyn, D.E., Dong, L., Dahdul, W.M., Stewart, C., Berger-Wolf, T., et al.: Bioclip: A vision foundation model for the tree of life. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19412–19424 (2024)
36. Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., Packer, C.: Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data* **2**(1), 1–14 (2015)
37. Valan, M., Picek, L.: Mastering large scale multi-label image recognition with high efficiency overcamera trap images. arXiv preprint arXiv:2008.07828 (2020)
38. Vidal, M., Wolf, N., Rosenberg, B., Harris, B.P., Mathis, A.: Perspectives on individual animal identification from biology and computer vision. *Integrative and comparative biology* **61**(3), 900–916 (2021)
39. Wightman, R.: Pytorch image models. <https://github.com/rwightman/pytorch-image-models> (2019). <https://doi.org/10.5281/zenodo.4414861>
40. Willi, M., Pitman, R.T., Cardoso, A.W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., Fortson, L.: Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution* **10**(1), 80–91 (2019)
41. Willi, M., Pitman, R.T., Cardoso, A.W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., Fortson, L.: Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution* **10**(1), 80–91 (2019). <https://doi.org/10.1111/2041-210X.13099>, <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13099>
42. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
43. Yu, X., Wang, J., Kays, R., Jansen, P.A., Wang, T., Huang, T.: Automated identification of animal species in camera trap images. *EURASIP Journal on Image and Video Processing* **2013**, 1–10 (2013)