Sieving Camera Trap Sequences in the Wild

Anoushka Banerjee, Dileep Aroor Dinesh and Arnav Bhavsar
MANAS Lab, SCEE, Indian Institute of Technology Mandi, Kamand, H.P., India

Keywords: Camera Trap, Empty Frames, Wildlife Detection, Domain Adaptation, Domain Generalization, Vision Transformer (ViT), Faster Region Based Convolution Networks (Faster R-CNN) and DEtection TRansformer (DETR).

Abstract: Camera trap sequences are a treasure trove for wildlife data. Camera traps are susceptible to false triggers caused by ground heat flux and wind leading to empty frames. Empty frames are also generated if the animal moves out of the camera field of view in between the firing of a shot. The time lost in manually sieving the surfeit empty frames restraint the camera trap data usage. Camouflage, occlusion, motion blur, poor illumination, and a small region of interest not only make wildlife subject detection a difficult task for human experts but also add to the challenge of sifting empty frames from animal containing frames. Thus, in this work, we attempt to automate empty frame removal and animal detection in camera trap sequences using deep learning algorithms such as vision transformer (ViT), faster region based convolution networks (Faster R-CNN), and DEtection TRansformer (DETR). Each biodiversity hotspot has its characteristic seasonal variations and flora and fauna distribution that juxtapose the need for domain generalization and adaptation in the leveraged deep learning algorithms. Therefore, we address the challenge of adapting our models to a few locations and generalising to the unseen location where training data is scarce.

1 INTRODUCTION

Ecological object detection has recently gained momentum due to the availability of high-tech non-intrusive and unobtrusive camera traps (Norouzzadeh et al., 2018) (Zhang et al., 2016) (Figueroa et al., 2014). Camera traps are being widely deployed to continuously observe natural habitats for months or even years, recording the rarest events occurring in nature (Norouzzadeh et al., 2019) (Swinnen et al., 2014) (Zhou, 2014). A colossal amount of data is generated in the process. Biologists and conservation activists need to sift the generated data for vital statistics on species locations, population sizes, interactions among species, behavioural activity patterns, and migrations.

Camera traps are heat triggered or motion sensed devices that passively record wildlife presence. Camera traps are prone to false triggers from moving vegetation, ground heat flux, passers-by, and wind (Beery et al., 2019). The camera trap can record an empty frame if the animal moves out of the field of view in between the trigger and the shot capture delay. Thus, more than 70% of camera trap images do not contain animal (Cunha et al., 2021) (Swanson et al., 2015) (Beery et al., 2019). Therefore, as a precursory to animal detection empty frames need to be removed.

The primary challenges in empty frame removal and animal detection in camera trap sequences are camouflage, motion blur, occlusion, poor and varying illumination, cropped out subject, subject too close or too far, varying animal poses and optical distortion due to fixed camera angles (Norouzzadeh et al., 2018) (Weinstein, 2018) (Beery et al., 2018). The exigence of empty frame segregation and animal detection in camera trap images can be attributed to the sparsity and sporadicity of subject content (Cunha et al., 2021). The subject is seldom in the centre and balanced with the background. Most images are difficult even for human eyes for identifying the presence of an animal. Frequently only a part of the animal is seen in low contrast frames due to dominant nocturnal wildlife and camouflage (Beery et al., 2018) (Norouzzadeh et al., 2018). The time lost in manual review and annotation is a bottleneck curtailing the use of camera traps for comprehensive large-scale environmental studies (Weinstein, 2018).

In our work, we elucidate a deep learning approach for automatic empty frame removal and detection and localisation of animals in their natural sur-

Banerjee, A., Dinesh, D. and Bhavsar, A.
Sieving Camera Trap Sequences in the Wild.
DOI: 10.5220/0010919000003122
In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022)*, pages 470-479
ISBN: 978-989-758-549-4; ISSN: 2184-4313

roundings. We use a contemporary object detection model; faster region convolutional network (Faster R-CNN) (Ren et al., 2015), a relatively recent transformer encoder-decoder based detection model; DEtection TRansformer (DETR) (Carion et al., 2020), and a solely attention based classification model; Vision Transformer (ViT) (Dosovitskiy et al., 2020). We perform our experimental studies on Caltech camera traps (CCT) dataset (Beery et al., 2018). The proposed approach can enable the extraction of valuable information from camera trap sequences speedily and eliminate human bias. Every biome has its distinctive vegetation, soil, climate, and wildlife. Therefore, we critically assess the prowess of the above mentioned deep learning algorithms on domain adaptability and generalisability in empty frame removal task.

In a nutshell, our contributions are as follows: (1) solve empty frame removal task in camera trap sequences and beat the state-of-the-art using deep learning algorithms such as ViT, Faster R-CNN and DETR, (2) animal detection in camera trap dataset and set a benchmark performance (3) demonstration of domain adaptability and generalisability trade-off of each model along with a critical assessment of the algorithms, (4) elucidating the prowess of the best performing models through attention maps, and (5) a two-stage pipeline for processing camera trap sequence to detect animals.

The rest of this paper is organized as follows: Section II briefly reviews related works on wildlife detection and camera trap sequences. Section III presents our proposed empty frame removal and animal detection approach, and Section IV describes the experiments, results, inferences and discusses the proposed pipeline. At the end Section V concludes the paper.

2 LITERATURE REVIEW

In the recent past, there has been a significant upsurge in research endeavours by the deep learning community to develop reliable animal detection algorithm. Animal detectors are not only useful for collecting ecosystem statistics but also for developing collision avoidance systems (Matuska et al., 2016). Matsuka et al.(Matuska et al., 2016) used scale invariant feature transform (SIFT) (Lindeberg, 2012) and speeded up robust features (SURF) (Bay et al., 2006) for local feature description. Then combined support vector machines (SVM) (Cortes and Vapnik, 1995) with radial base function and bag of visual keypoint method for background subtraction and region of interest (ROI) detection. After localising the ROI, continuously adaptive mean shift (CAMShift) algorithm was applied to find optimal object size, position, and orientation. But this methodology has the limitation that if the animal is static for too long, the detector will produce a false negative. As discussed in (Emami and Fathy, 2011) (Hidayatullah and Konik, 2011), CAMShift has reliable performance with single hue object tracking and in scenarios where object colour has high contrast with the background colour. But fails miserably with multi-hue object tracking and cases where the object blends with the background. Over a million years of evolution, organisms adjusted to their surroundings by camouflaging. Blending with the backdrop, birds and animals are the lion's share successful in survival and procreation. Therefore, the deployment of CAMShift algorithm in the wildlife detection system does not seem to be a wise decision.

Figueroa et al. (Figueroa et al., 2014) and Swinnen et al.(Swinnen et al., 2014) relied on handcrafted feature to detect animals. Low-level pixel changes were used as a delimiter between frames containing animals and not containing animals. These techniques have a strong correlation with environmental circumstances for accurate detection. Natural habitats are prone to seasonal changes, day-night lighting variation, mist, and haze. For the current task at hand, we should work out a modus operandi which ideally has a low correlation with seasonal and day-to-day changes in environmental conditions. Furthermore, all the enlisted algorithms (Figueroa et al., 2014) (Swinnen et al., 2014) (Emami and Fathy, 2011) (Matuska et al., 2016) work best on medium to large mammals. The performance significantly degrades when ROI is smaller.

In (Zhou, 2014) the effectiveness of diverse feature generation algorithms for animal detection such as local binary pattern (LBP) (Guo et al., 2010) and histogram of gradients (HOG) (Dalal and Triggs, 2005) is investigated. LBP introduced by Ojala et al. (Ojala et al., 1996) was designed for monochrome static images and has the limitation that the feature size increases with the number of neighbours. Torralba and Olivia (Torralba and Oliva, 2003) studied the statistical properties of natural images. Their results concluded that without segmentation or grouping stages low-level features can be used for object localisation. Zhang et al. (Zhang et al., 2016) constructed iterative embedded graph cut (IEC) method for region proposal. This technique is not robust to background variations and produces false positives due to moving clouds, waving leaves, and shadows. Although after several cross-frame fusion better results may be obtained.

Ensemble of CNN archetypes such as VGG (Simonyan and Zisserman, 2014), AlexNet (Krizhevsky

et al., 2012), ResNet (He et al., 2016), GoogLeNet (Szegedy et al., 2015) and NiN (Lin et al., 2013) was used for automatic counting and detecting animals in Snapshot Serengeti dataset (Norouzzadeh et al., 2018). As Bounding Box is not available with the Snapshot Serengeti dataset, counting was treated as a classification problem with ±1 error tolerance.

More recently, approaches predicated on deep convolutional neural networks (DCNN) such as Faster R-CNN (Ren et al., 2015) and its predecessors Fast R-CNN (Girshick, 2015) and R-CNN (Girshick et al., 2014) are attaining the state-of-the-art performance. Customarily, these algorithms are comprised of two major segments. The first segment generates object region proposals at different scales and locations and the second segment performs foreground identification to tell if the region proposal truly contains an object or not. Beery et. al (Beery et al., 2018), exploited the theme of generalization across datasets, backdrops, and locations using Faster R-CNN with ResNet and Inception backbone. Schneider et. al (Schneider et al., 2018) demonstrated the detection capabilities of Faster R-CNN and YOLOv2.0 on Reconyx camera trap and Snapshot Serengeti datasets. Since natural scenes are cluttered, these methods generate a large number of region proposals. They ignore the unique spatio-temporal features in the context of animal detection from camera trap images (Zhang et al., 2016). Faster R-CNN algorithm needs many passes through a single image to detect all the objects.

Few works address the problem of automating empty frame removal. Cunha et al. (Cunha et al., 2021) examined the trade-off between precision and inference latency on edge devices in empty frame removal task. The use of attention mechanisms in object detection is recently gaining prevalence. In the recent past, a few approaches amalgamated convolutional features and attention mechanisms.

3 PROPOSED STUDY, DATASET AND DATA SPLITS

A very demanding and tiring aspect of unsheathing vital statistics from camera trap sequences is empty frame removal followed by detecting animals in the sea of natural background clutter. Thus, to reduce the time lost in manual review and eliminate human bias we propose to solve the empty frame removal and animal detection task in camera trap sequences. We leverage ViT, DETR, and Faster R-CNN architecture for this purpose. The contributors of the CCT dataset used Faster R-CNN for their experimental analysis and had set a benchmark (Beery et al., 2018). Faster R-CNN from its inception is widely used across manifold applications and domains for object detection. Thus, as an obligatory choice, we employ Faster R-CNN for empty frame removal and animal detection task for comparisons.

But, the CCT dataset is laden with challenges such as the dominance of low-contrast images, a small range of tones, varying subject size, pose variations, and occlusions. Most existing deep learning approaches use state-of-the-art convolutional neural network (CNN) backbones for object detection such as faster region based CNN (Faster R-CNN) (Ren et al., 2015), single shot detector (SSD) (Liu et al., 2016) and you look only once (YOLO) detection algorithms (Redmon and Farhadi, 2017). But convolution operation has the limitation that all image pixels are treated equally (Wu et al., 2020). Semantic concepts with long-range dependence are often lost. Edges and corners are easily captured by convolution filters but sparse high-level concepts escape the sieve (Wu et al., 2020). Therefore, for camera sequences, we need a paradigm of detectors that alleviates the above stated impediments.

Animals in the wild are often occluded which needs that despite discontinuity in animal image pixels the distant but related animal features be interweaved. Therefore, we choose to experiment with DETR; a relatively recent paradigm of detectors that reason about the relations of the animals with the global image content. DETR based upon encoder-decoder mechanism employing multiple attention heads performs global reasoning. Global reasoning aids the model in making decisions about occluded and obscured object parts. Occlusion and low illumination being the most common ailment of the dataset, transformer architecture should have a clear advantage here.

Transformers lack inductive biases such as locality and translation equivariance inherent to CNNs (Dosovitskiy et al., 2020). Intuitively, the lack of locally restrictive receptive field and translation invariance in transformers can be desirable for empty frames segregation tasks. Vision Transformer (ViT) (Dosovitskiy et al., 2020) is a pure transformer model. It inputs the image as a sequence of patches in the form of tokens. The image patches tease out patterns efficiently. Therefore, inspired by ViTs success in vision and to pose empty frame segregation purely as a classification problem we roped in ViT.

3.1 Caltech Camera Traps (CCT) Dataset

There are 243,100 frames from 140 locations in the Southwestern United States. The object categories seen in this dataset are bobcat, opossum, coyote, raccoon, bird, dog, cat, squirrel, rabbit, skunk, lizard, rodent, badger, deer, cow, fox, car, badger, fox, pig, mountain lion, bat, and insect. Approximately 70% of frames are blank. CCT dataset has a subset dataset called Caltech Camera Traps-20 (CCT-20) (Beery et al., 2018). It contains 57, 868 images across randomly chosen 20 locations from 140 locations of CCT dataset.

3.1.1 Data Split for Empty Frame Removal

The image distribution for any camera trap sequence in general and CCT dataset, in particular, is skewed towards empty frames as seen in Figure 1. To eliminate bias we balance the number of empty frames and animal frames at each location. For training data, we choose empty and animal images from the 20 locations same as in CCT-20 dataset. For testing we create two sets; (1) 'cis': the test data from locations seen by the model during training and (2) 'trans': the test data from locations unseen by the model during training as in (Beery et al., 2018). To balance the number of animal and empty frames at each training location we take the minimum of the maximum number of animal and empty frames at each location. We extract 8,028 images containing equal numbers of animal and empty frames. We use 70% of the extracted images for training and 30% for testing the model. Therefore, we have 5,638 training examples and 2,390 testing examples as in Figure 2 for seen locations. To analyze domain adaptation vs. domain generalization we pulled out 1,195 empty and animal frames each from locations other than 20 training locations. These 2,390 images have background characteristics different from locations seen during training.

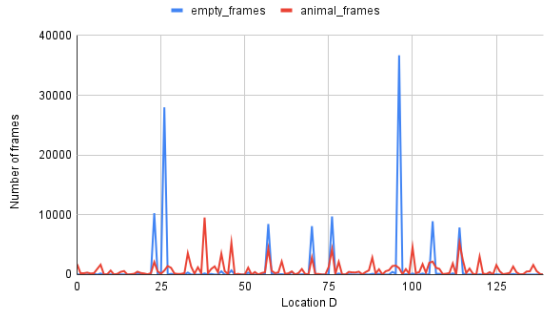

Figure 1: Number of empty and animal frames vs. locations in entire dataset.

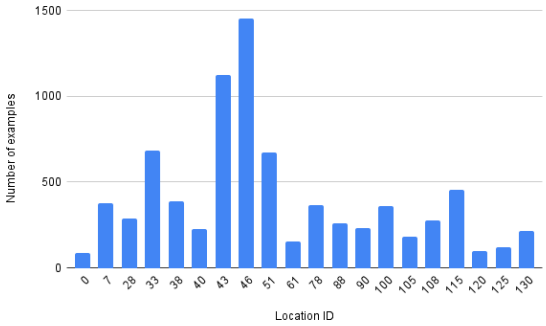

Figure 2: Number of examples vs. locations in dataset for empty frame removal experiments.

3.1.2 Data Split for Animal Detection

Every location in CCT dataset has its characteristic flora and fauna, seasonal variations, frequency of occurrence of species, and day and night duration. Therefore, to critically assess the robustness and accuracy of the leveraged algorithms in the light of domain generalisation we use completely disjoint location sets for train and test. From 140 locations in the CCT dataset, we use 100 locations 'cis' for model training and 40 unseen 'trans' locations for test. We have 37,356 training images with 39,361 annotations and 24,489 test images with 25,571 annotations.

4 EXPERIMENTS, RESULTS AND DISCUSSIONS

Our experimental analyses are bifurcated into; (1) empty frame removal experiments and (2) animal detection in camera trap sequences.

4.1 Empty Frame Removal

We use ViT, Faster R-CNN, and DETR for segregating empty frames from animal frames. We pose empty frame segregation as a binary classification problem. The two classes are frames containing animal (animal frames) and frames without animals (empty frames). We use percentage accuracy as the evaluation metric. We train the models on 'cis' locations and test the prowess on both 'cis' and 'trans' test sets.

4.1.1 Experiments with Vision Transformer(ViT)

Empty frames are natural backdrops without any object. The non-empty frames are natural backdrops with an animal. Therefore, to model empty frame

segregation purely as a classification task we leverage ViT. We train two models: (1) The first model adapts better i.e. the best model on 'cis' locations. In this case, the model is trained till the test accuracy on the 'cis' location is maximum. (2) The second model generalises better i.e. the best model on 'trans' location. In this case, the model is trained till the best accuracy on 'trans' location is obtained. The ViT models are trained using AdamW (Loshchilov and Hutter, 2018) with a weight decay of 0.0001 and an initial learning rate of 0.001. We employ random horizontal flip, random rotation, and random zoom with a factor of 0.2 as data augmentation techniques.

Results from Best Model on 'cis' Locations: We obtain an overall accuracy of 87.28% on 'cis' and 66.28% on 'trans' (Table 1). This can be accredited to the generalization gap. On 'cis', the performance for empty frames is better and on the flip side for 'trans' the performance on animal frames is better. In camera trap sequences the backdrop for both frames containing animals and empty frames for a particular location is the same. Therefore, for 'cis' locations the model has a bias towards empty frames. In contrast, for 'trans' locations some species are common as in 'cis' but the backdrop is different resulting in better performance for animal frames.

Table 1: Accuracy in % from ViT best model on 'cis' locations.

	Animal	Empty	Total
'cis'	84.60	89.96	87.28
'trans'	69.29	63.26	66.28

Results from Best Model on 'trans' Locations: We obtain an overall accuracy of 67.15% on 'cis' locations and 72.93% on 'trans' locations (refer to Table 2). Astoundingly, on 'cis' locations the animal performance improved from 84.60% to 93.72% and empty frames performance dwindled from 89.96% to 40.59% in comparison to 'cis' best model (from Table 1 and Table 2). Every location has its characteristic vegetation and camera field of view. Therefore, in the case of 'trans' locations the empty frames are completely unseen. Hence, 'trans' best model has very high accuracy for animal frames but much lower empty frame performance.

Table 2: Accuracy in % from ViT best model on 'trans' locations.

	Animal	Empty	Total
'cis'	93.72	40.59	67.15
'trans'	88.37	57.49	72.93

4.1.2 Experiments with Faster R-CNN

We wield pretrained Faster R-CNN with ResNet-101 backbone available in Detectron2 codebase (Wu et al., 2019). The empty frames are annotated with a null tensor; $[0,0,0,0]$ bounding boxes; to indicate 0 dimensions and null area for no subject content. An empty frame is considered correctly classified if the confidence threshold and predicted bounding boxes are null tensors. We use SGD with a momentum of 0.9. Beginning at 0.001 we decay the learning rate by 0.05 after every 1000 epochs.

Results from Faster R-CNN: We obtain an overall accuracy of 65.35% from 'cis' locations and an overall accuracy of 64.27% from 'trans' locations (Table 3). Faster R-CNN is an object detection algorithm. Therefore, the accuracy for detecting animals is much higher for both 'cis' and 'trans' location sets. There are many false positives in this case; defeating the objective of empty frame removal.

Table 3: Accuracy in % from Faster R-CNN.

	Animal	Empty	Total
'cis'	98.74	31.96	65.35
'trans'	95.06	33.47	64.27

4.1.3 Experiments with DETR

For segregating empty frames from animal frames using DETR, we pose empty frame removal as a by-product of the detection problem. The DETR model is finetuned with AdamW (Loshchilov and Hutter, 2018) with a learning rate at 0.0001. Random horizontal flips is used as a data augmentation technique. We train DETR on animal instances from 'cis' and filter out empty frames using confidence thresholding. In the training set, all the animal instances are associated with a confidence score greater than 0.95 (refer to Figure 3). Therefore, we choose 0.95 as the confidence threshold for segregating empty frames. During testing, a frame is considered empty if the confidence score associated with the frame for animal presence is less than 0.95.

Results from DETR: From DETR we obtain an overall accuracy of 80% on 'cis' locations and overall accuracy of 76.57% on 'trans' locations (see Table 4). The accuracy obtained on identifying frames containing animals is appreciable 98.49% on 'cis' locations and 90.71% for 'trans' locations. We infer that DETR has the least generalisation gap due to the encoder-decoder attention mechanism.

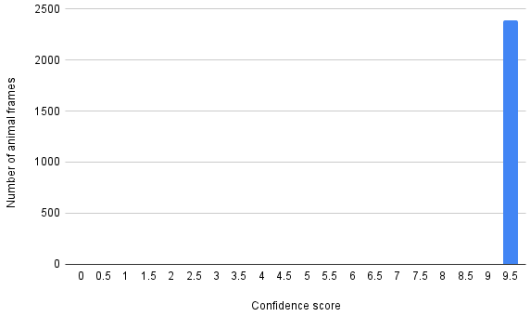

Figure 3: Number of animal frames in training set vs. confidence score from DETR.

Table 4: Accuracy in % from DETR.

	Animal	Empty	Total
'cis'	98.49	61.51	80.00
'trans'	90.71	62.43	76.57

The inferences drawn from the above studies are given in the next subsection. These inferences will be used to narrow down upon the most suitable deep learning algorithm for empty frame removal.

4.1.4 ViT, Faster R-CNN or DETR?

'cis' Location Performance: Among the three deep learning techniques, the highest overall accuracy on the 'cis' locations (87.28%) is given by ViT (best model on 'cis' locations).

Faster R-CNN, ViT (best model on 'trans' locations), and DETR for animal containing frame have accuracy more than 90% but for empty frame, the accuracy is 40.59%, 31.96% and 61.51% respectively. These models will detect animals with high accuracy but a large mass of empty frames will be wrongly identified as animal containing frames. Thus the primary motive of removing empty frames will not be solved. On the contrary, an important consideration for infrequently encountered and rare species is that we want to retain any frame containing even slight traces of a species. Hence, in such a scenario we might relax empty frame removal if the cost is losing frames containing uncommon species. DETR and Faster R-CNN bequeath us with nearly the same accuracy on animal frames (98.49% and 98.74% respectively). But DETR has an upper hand on empty frame performance by nearly 30% in comparison to Faster R-CNN. Therefore, while dealing with rare species we should employ DETR. Most camera trap sequences produce a colossal amount of data with more than 70% empty frames. Therefore the sweet spot between reducing the burden of manual empty frame removal and retaining rare species lies in the usage of ViT('cis' best model.)

'trans' Location Performance: The highest overall accuracy on 'trans' locations 76.57% is given by DETR. DETR generalises well to new locations as it focuses on animal presence. From the DETR self-attention maps we observe that DETR coalesces distant semantic concepts well as all the background pixels are interweaved and differentiated from object pixels (Figure 6). Therefore, DETR provides nearly the same accuracy on empty frame for both 'cis' and 'trans' locations (61.51% and 62.43% respectively).

At the same time, if we want to retain the maximum number of frames containing animals belonging to extremely rare species Faster R-CNN is the most suitable model.

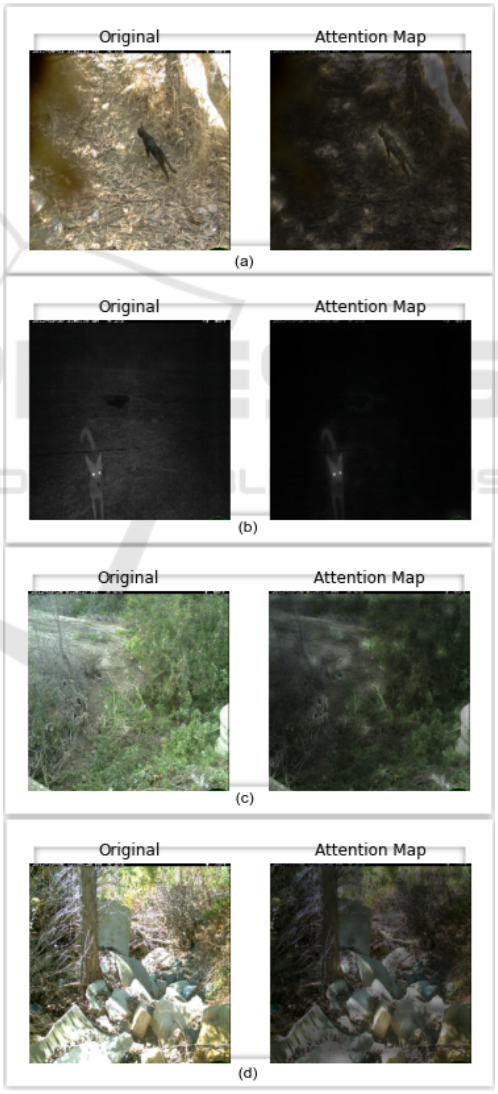

Figure 4: Attention map for: (a) frame containing animal at location 88, (b) frame containing animal at location 7, (c) empty frame at location 130 and (d) empty frame at location 40.

ViT- Appraisal through Attention Map: Despite the multitude of challenges, including varying poses, camouflage, occlusion due to tree branches and leaves, low contrast, and poor illumination ViT focuses its attention solely upon the animal (refer to Figure 4 (a) and (b)). From Figure 4 (b) we see that even though animal is not salient due to low contrast, no tonal gradient, and low illumination the attention rightly focuses itself on the animal. In spite of background clutter, deceiving animal-like objects, and presence of colour tones similar to that of animals in empty frames the attention is not localised, it is scattered and dispersed throughout the frame as seen in Figure 4 (c) and (d). Therefore, through visual scrutiny of attention maps, we see that ViT rightly focuses on the animals.

In the next section, we present results on detecting (localising) animals in camera trap sequences from frames containing animals.

4.2 Animal Detection

For detecting animals in camera trap sequences we use Faster R-CNN and DETR. The evaluation metric used for the animal detection task is COCO Average Precision (AP). COCO AP $0.5-0.95$ denotes the average over multiple IoU (Intersection over Union) threshold from 0.5 to 0.95 with a step size of 0.05. The detection above the IoU threshold is taken as the correct detection. COCO AP 0.5 is Average Precision above 0.5 IoU threshold.

4.2.1 Animal Detection Results

From Table 5 it is observed that DETR outperforms Faster R-CNN in the animal detection task. This observation could be accredited to the knitting of sparse-high level semantics together in camera trap images done by the fusion of convolutional backbone, attention mechanism, and encoder-decoder architecture in DETR. Hence, for further scrutiny, we employ DETR. The reason for the better performance of DETR is explained as beneath.

Table 5: Detection performance using DETR and Faster R-CNN in AP(COCO AP 0.5-0.95).

	Faster R-CNN	DETR
AP 0.5-0.95	52.1	53.4

Sequence Analysis of Camera Trap Images with Results from DETR: Since camera trap images are a sequence of images, not all frames captured are equal. After a trigger to shoot images it is an optimistic conjecture to assume that at least one frame can capture the animal. Therefore, as in (Beery et al., 2018), we exploit frame information in two ways:

1. **Most Confident:** From a sequence of images, if the most confident detection from all the frames in a sequence has an IoU greater than 0.5 we consider the animal has been correctly located.
2. **Oracle:** From a sequence of images, if the highest IoU between the predicted bounding box and ground truth bounding box (amongst all the frames in a sequence) is greater than 0.5 we consider the animal has been correctly located in that sequence.

For sequence analysis, we remove the images with multiple animals to have a clear case for choosing the highest IoU and highest confidence amongst a sequence of frames.

Treating all the frames equally i.e. without sequence information we obtain an average precision of 89.2 (COCO AP 0.5) (refer to Table 6). With sequence information *Most Confident* and *Oracle* we wield average precison of 91.4 and 94 (COCO AP 0.5) respectively. The frame sequences in this dataset vary from 1-5 in length. Hence, it is appropriate to evaluate the performance using sequence information. 94 COCO AP 0.5 points seem to be an upstanding score considering the preposterous challenges of the dataset; camouflage, motion, blur, occlusion, poor illumination, negligible tonal range, cropped out subject, subject too close or too far, varying animal poses, and optical distortion due to fixed camera angles.

Table 6: Sequence analysis of camera trap images with results from DETR.

	No sequence information	Most confident	Oracle
AP 0.5-0.95	55.4	61.2	68.5
AP 0.5	89.2	91.4	94

DETR Detection Performance Scrutiny:. To appraise the performance of DETR on CCT sequence 1000 images are randomly selected for visual scrutiny. Near flawless detection is observed in most of the daytime images (image(1) in Figure 5) and in fair proportion of low-light images (image(2) and image(3) in Figure 5). Despite the rationale that luminance and colour gradient is pivotal for shape and depth inference by any algorithm, DETR performs well on a fair share of images that have nearly flat or very weak luminance gradient and negligible range of tones.

Many failure cases involve deceiving animal-like background clutter (image(4)), or very low visibility

Figure 5: DETR detection output on CCT dataset from left to right and top to bottom: (1) IoU equal to 0.99 (2) IoU equal to 0.79 (3) IoU equal to 0.63 (4) False positive; IoU equal to 0.39 (5) False negative detection (6) False positive detection (The bounding boxes in red are ground truth and blue are predicted).

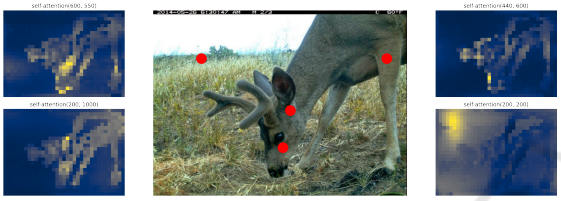

Figure 6: Visualization of DETR encoder self-attention weights for a CCT image; first column top and bottom: self attention map for reference point belonging to the object; second column: original image; third column top and bottom: self attention map for reference point belonging to the object and belonging to background respectively.

either due to low-light (image(5)), or extreme camouflage (image(6)). Indeed in such cases, it is difficult to see the animal, and there seems to be only a possibility but no guarantee of animal presence.

In Figure 6, we visualise DETR encoder self-attention using four reference points. Self attention for a sample point indicates the likelihood of remaining points being positively related to it. By visualising how the model encodes the object, we gain intuition whether the model is accurately knitting context semantics while maintaining spatially-distant concepts. So, in Figure 6 we see that for the three reference points belonging to the object, the model assigns maximum weight to pixels belonging to the subject. Even if there are three different spatially apart reference points belonging to the object, the self-attention matrix visualised is nearly the same. Thus, giving less weight to background pixels, the model filters out irrelevant parts of the image endowing itself with the ability to make precise judgements. Also, for a background reference point, the entire background is weighted higher than object pixels. This gives us the insight that indeed DETR encoder is maintaining long-range dependencies (by weaving the background pixels together throughout the image). The ability to capture dependencies beyond the limited receptive

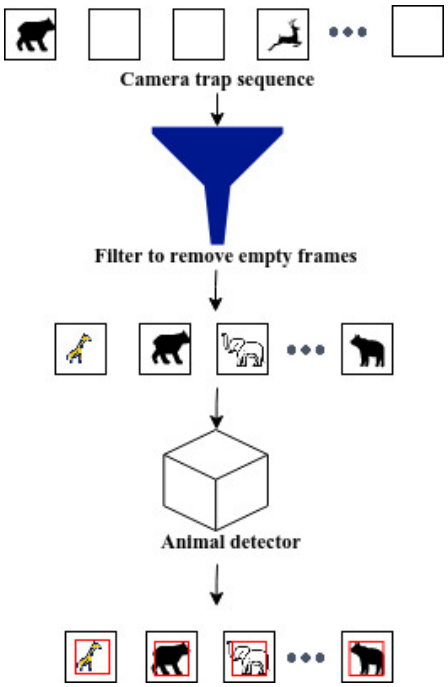

Figure 7: Proposed pipeline for empty frame removal and animal detection.

field of a convolutional filter is the key to better performance of DETR than Faster R-CNN on the CCT dataset. This further allows the model to gain wider intuition for detecting obscured animal parts.

Based on the results from section 4 (empty frame removal and empty frame detection) we propose an end-to-end pipeline for animal detection from camera trap sequences.

4.3 Discussion on a Proposed Pipeline for Camera Trap Image Analysis

Encouraged by our results, we propose an end-to-end pipeline for empty frame removal and animal detection task in camera trap sequences (refer to Figure 7). Given a camera trap sequence, the proposed pipeline removes the empty frames using ViT (best model on 'cis') and then locates animals with a bounding box using DETR. It is observed that in CCT dataset 70% of the frames are empty. The proposed pipeline is applied to the entire data and 99.4% of the total empty frames are discarded. Only 0.6% of the total number of empty frames remain along with frames containing animals. At the same time, we have lost 0.2% of the frames that contain animals. In the next stage of the pipeline, it is seen that 87.29% of animals are detected with IoU greater than 0.5. IoU greater than 0.5 is a widely used threshold for object detection tasks. We observe that 98.56% of animal frames are

detected with IoU greater than 0.3. With IoU above 0.3, the task of locating animals becomes very easy in extremely low light and low contrast images.

5 CONCLUSION

In this work, we address the empty frame removal problem and the animal detection challenge in camera trap sequences. In tandem, we investigate the applicability of ViT, DETR, and Faster R-CNN for this task. Our experiments reaffirm the generalisation gap in the context of unseen test data. We culminate our experimental study with proposal of a two-stage pipeline for mining vital statistics from camera trap sequences. In the first stage we filter out empty frames and in the second stage, we perform wildlife detection and localisation. Balancing the trade-off between retaining all frames containing animals and filtering out all empty frames we adopt ViT(best model on 'cis') for removing empty frames and DETR for detecting animals. Despite heavy background clutter, camouflage, size and pose variations, occlusion, progressive illumination changes from day to night, and seasonal variations in flora and fauna in camera trap data we obtain a competitive accuracy. We shall extend our work to make the empty frame removal and animal detection pipeline even more robust, especially under extreme low-light and low-contrast conditions. Hence, develop practically deployable wildlife detection systems. Further, we plan to incorporate open set recognition, zero-shot learning, and few-shot learning for generalising to unseen locations.

ACKNOWLEDGEMENTS

This work is partially supported by National Mission for Himalayan Studies (NMHS) grant GBPNI/NMHS-2019-20/SG/314

REFERENCES

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer.

Beery, S., Van Horn, G., Mac Aodha, O., and Perona, P. (2019). The iwildcam 2018 challenge dataset. *arXiv preprint arXiv:1904.05986*.

Beery, S., Van Horn, G., and Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cunha, F., dos Santos, E. M., Barreto, R., and Colonna, J. G. (2021). Filtering empty camera trap images in embedded systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2438–2446.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Emami, E. and Fathy, M. (2011). Object tracking using improved camshift algorithm combined with motion segmentation. pages 1–4.

Figueroa, K., Camarena-Ibarrola, A., García, J., and Villela, H. T. (2014). Fast automatic detection of wildlife in images from trap cameras. In *Iberoamerican Congress on Pattern Recognition*, pages 940–947. Springer.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Guo, Z., Zhang, L., and Zhang, D. (2010). A completed modeling of local binary pattern operator for texture classification. *IEEE transactions on image processing*, 19(6):1657–1663.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hidayatullah, P. and Konik, H. (2011). Camshift improvement on multi-hue and multi-object tracking. In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, pages 1–6.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Lindeberg, T. (2012). Scale invariant feature transform.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

Loshchilov, I. and Hutter, F. (2018). Fixing weight decay regularization in adam.

Matuska, S., Hudec, R., Kamencay, P., and Trnovszky, T. (01 May. 2016). A video camera road sign system of the early warning from collision with the wild animals. *Civil and Environmental Engineering*, 12(1):42 – 46.

Norouzzadeh, M. S., Morris, D., Beery, S., Joshi, N., Jojic, N., and Clune, J. (2019). A deep active learning system for species identification and counting in camera trap images. *ArXiv*, abs/1910.09716.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., and Clune, J. (2018). 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.

Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51 – 59.

Redmon, J. and Farhadi, A. (2017). Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*.

Schneider, S., Taylor, G. W., and Kremer, S. (2018). Deep learning object detection methods for ecological camera trap data. In *2018 15th Conference on computer and robot vision (CRV)*, pages 321–328. IEEE.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., and Packer, C. (2015). Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data*, 2(1):1–14.

Swinnen, K. R., Reijniers, J., Breno, M., and Leirs, H. (2014). A novel method to reduce time investment when processing videos from camera trap studies. *PloS one*, 9(6):e98881.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Torralba, A. and Oliva, A. (2003). Statistics of natural images categories. *Network (Bristol, England)*, 14:391–412.

Weinstein, B. G. (2018). A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3):533–545.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Tomizuka, M., Keutzer, K., and Vajda, P. (2020). Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2.

Zhang, Z., He, Z., Cao, G., and Cao, W. (2016). Animal detection from highly cluttered natural scenes using spatiotemporal object region proposals and patch verification. *IEEE Transactions on Multimedia*, 18(10):2079–2092.

Zhou, D. (2014). *Real-time animal detection system for intelligent vehicles*. PhD thesis, Université d'Ottawa/University of Ottawa.