ECO SCAN - AI - POWRED ANIMAL RECOGNITION AND SPECIES CATEGORIZATION

Saravanan T[1], Sowmiya S[2], Kishore N S[3], Nikitha M[4]

[1] Student, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India
[2] Student, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India
[3] Student, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India
[4]Faculty, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India

ABSTRACT:

It is essential to identify and categories animal species in order to evaluate their long-term survival and the potential effects of our actions on them. This procedure also helps identify predatory and non-predatory species, both of which represent serious risks to both people and the environment. Additionally, it helps to lessen traffic accidents in several areas where contacts with animals on the road have caused countless car accidents. However, obstacles like size differences and divergent behaviour between species make it difficult to identify and categories animal species. In order to build an integrated system that successfully addresses these issues, the novel two-stage network and modified multi-scale attention mechanism presented in this research are used. We adopt a pyramid design with lateral connections at the regional proposal stage to increase the sensitivity of semantic properties for smaller objects. In order to improve functional transmission and multiplex it across the classification step, we also use a densely linked convolutional network, which leads to more accurate classification with fewer parameters. Our experiment showcases the autonomous data extraction capabilities of deep neural networks, a cutting-edge type of artificial intelligence. To fully utilize the potential of these technologies, the ultimate goal is to train neural networks for autonomous animal identification and recognition.

Keywords: *Animal detection, Feature learning, Image modalities, Deep neural network, camera trap images.*

1. INTRODUCTION:

Identifying and identifying animal species is essential for tackling problems like human-wildlife conflicts and wildlife-related road accidents that result in fatalities and injuries (Nowak et al., 5). Animal attacks, which have resulted in numerous fatalities and injuries in humans, occur at various rates according on location. For instance, it is estimated that there are two million animal assaults on people each year in the US (Warrell, 6). Between 1990 and 2005, at least 563 villages reported having their residents attacked by animals, according to Tanzanian and American scientists. Predatory animals like tigers and lions are recognised to provide serious threats; more human fatalities have been caused by tigers than by any other species of their kind (Nowak et al., 5). However, the full magnitude of animal-related mortality is hidden by the absence of thorough records among governments. In order to reduce these hazards, avoid animal-vehicle accidents, and discourage theft, it is essential to create effective procedures for animal detection, classification, and monitoring. Animal attacks frequently happen at night because of hunger, as animals travel in search of food. The focus of these studies is object detection, a fast developing area of computer vision, with deep learning methods like CNNs exhibiting excellent performance in visual comprehension. Due to their great precision, two-stage detectors such Faster R-CNN, R-FCN, FPN, and YOLOv5 (Birds class, 7) have drawn a lot of interest. Anchor size issues still exist and have an impact on detecting precision.

The use of animal detection in the field of computer vision is essential for resolving a variety of issues, such as wildlife accidents and the preservation of endangered species (Birds class, 7). Animal identification presents special difficulties, mostly because different species can differ from one another in terms of structure, colour, and appearance (Birds class, 7). Animal identification is also impacted by variations in lighting and direction. Convolutional neural networks (CNNs) with fewer parameters and connections are needed for these challenges, which call for specialised models with strong learning skills to recognise various animal breeds in still images (Birds class, 7).

Along with the Region Proposal Network (RPN) to handle small animal species, attention methods in object detection and classification frameworks, including intricate and soft attention, have attracted interest. Small animal detection methods use image magnification and high-resolution detection maps, but multi-level representation network modifications improve model performance. However, the development of real-time applications poses computational difficulties in efficiently resolving these problems (Birds class, 7).

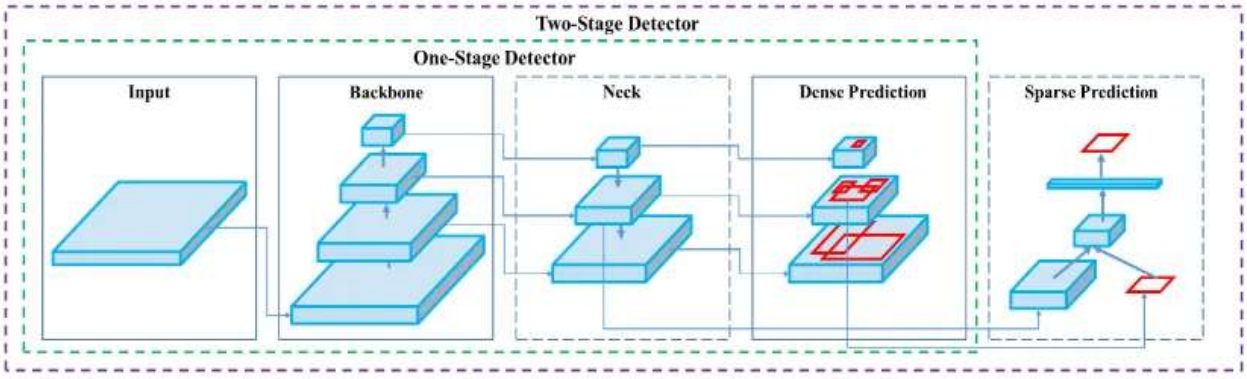

Figure 1 Object Detection

Method	Backbone	Size	FPS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOv4: Optimal Speed and Accuracy of Object Detection									
YOLOv4	CSPDarknet-53	416	38 (M)	41.2%	62.8%	44.3%	20.4%	44.4%	56.0%
YOLOv4	CSPDarknet-53	512	31 (M)	**43.0%**	**64.9%**	**46.5%**	**24.3%**	**46.1%**	**55.2%**
YOLOv4	CSPDarknet-53	608	23 (M)	43.5%	65.7%	47.3%	26.7%	46.7%	53.3%
Learning Rich Features at High-Speed for Single-Shot Object Detection [84]									
LRF	VGG-16	300	76.9 (M)	32.0%	51.5%	33.8%	12.6%	34.9%	47.0%
LRF	ResNet-101	300	52.6 (M)	34.3%	54.1%	36.6%	13.2%	38.2%	50.7%
LRF	VGG-16	512	38.5 (M)	36.2%	56.6%	38.7%	19.0%	39.9%	48.8%
LRF	ResNet-101	512	31.3 (M)	37.3%	58.5%	39.7%	19.7%	42.8%	50.1%
Receptive Field Block Net for Accurate and Fast Object Detection [47]									
RFBNet	VGG-16	300	66.7 (M)	30.3%	49.3%	31.8%	11.8%	31.9%	45.9%
RFBNet	VGG-16	512	33.3 (M)	33.8%	54.2%	35.9%	16.2%	37.1%	47.4%
RFBNet-E	VGG-16	512	30.3 (M)	34.4%	55.7%	36.4%	17.6%	37.0%	47.6%

Figure 2 Comparision of various Detection Methodology

1.1 DEEP LEARNING AND IMAGE CLASSIFICATION:

Mastering the art of mapping input data to desired output categories through the use of specialised neural network topologies is the primary goal in the field of deep learning, particularly within the domain of supervised learning (Goodfellow et al., 2016). The main objective of image classification is to develop a deep learning system that can analyse and classify photos into predetermined groups, including different animal species. Convolutional neural networks (CNNs) have dramatically increased in popularity in recent years, with notable challenges like the ImageNet Large Scale Visual

Recognition Challenges (ILSVRC) serving as notable examples of their dominance (Krizhevsky, Sutskever, & Hinton, 2012; Russakovsky et al., 2015).

LeCun et al. first described CNNs in 1989. A convolutional section is intended to extract localised features from images, and a fully connected segment is in charge of mapping these features to the desired output categories (LeCun et al., 1989). CNNs, in contrast to older methods, do not require manually created features. Instead, they learn spatial features on their own by automatically changing the parameters (weights) of the model while it is being trained. This is done by propagating errors from the output layer back to the input. The architecture of a CNN is determined by the specific arrangement of the processes carried out on the data. The architecture of a CNN is schematically shown in Figure 3, with the layer's main unit, a layer, which includes filters performing convolutions on the input data to identify spatial patterns, incorporating activation functions, and carrying out pooling (sub-sampling) operations, being highlighted. The subsequent layer is normally sent the smaller feature maps that each layer typically produces. The successive arrangement of several such layers makes it possible to extract detailed features. The number of layers that make up a neural network's design determines its depth, which represents the core of deep learning—neural networks with many of layers—according to He et al. (2015).

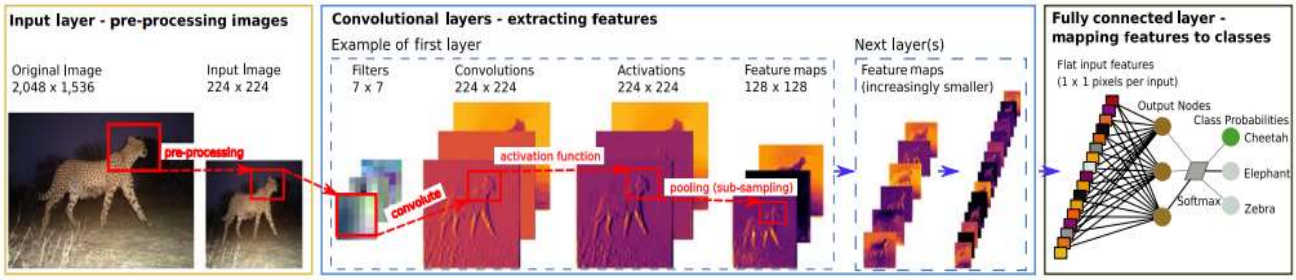

Figure 3 Schematic illustration of CNN architecture

2. METHODOLOGY

The training and testing phases of our classifier's operation are separate. Visual exemplars are a set of images used throughout the training process. The test image, which is supplied as input to the classifier in the following testing phase, is a recently acquired image. The classifier then assigns the test image to the most suitable class using the knowledge gained during training.

A. *Receiving the input image:*

An image is obtained by the system's attached camera in the system as it is now being imagined. The initial input is this taken test image, which is then transformed into a binary pattern. Then, using a dataset of previously labelled images, the test image's distinguishing traits are contrasted with those from the labelled images. This kind of comparison helps to identify the particular animal species included in the photograph.

B. *Feature Extraction*:

A condensed set of features can be produced by processing the input test image. These chosen features could include important information from the input data, making it possible to complete the required task with this condensed dataset as opposed to the original, unaltered data. Fixed features are directly taken from photographs, also referred to as human-crafted features. Deep neural networks, as opposed to features created by humans, are able to recognise elements within images and create several levels of representation, with the higher-level features encapsulating more abstract parts of the data.

C. *Identifying the species present in an image*:

The output layer calculates the probabilities associated with the existence of the detected animal in the image and assigns it to one of the possible classes in the context of species categorization. Although providing such an outcome could considerably minimise the amount of human labour needed for precise species identification, this hypothesis still has to be verified by humans because they have the skills and knowledge to do so.

3. LITERATURE SURVEY:

An substantial amount of human preprocessing was required in the early studies on automated animal identification, which mostly focused on matching species-specific patterns in photos. The achieved accuracy, which was 82% according to Yu et al. (2013), was less than the 96.6% reported by Swanson et al. (2016) for human-level accuracy. With some involving manual preprocessing (Gomez Villa et al., 2017) or more complex pipelines with automatic preprocessing (Giraldo-Zuluaga, Salazar, Gomez, & Diaz-Pulido, 2017), recent studies using Convolutional Neural Networks (CNNs) for automatic animal species identification have reported accuracies around 90%. Norouzzadeh et al. (2018)'s most recent developments achieved accuracy levels of 93.8%, matching human accuracy on more than 99% of photos.

In comparison to earlier studies, our study aims to deploy and validate CNNs across a wider range of camera trap datasets. The Snapshot Serengeti dataset, which has 3.2 million photos, was used by Norouzzadeh et al. (2018) to illustrate outstanding results, however most camera trap datasets that have been found on Zooniverse are less in size. Large datasets are frequently necessary for effective image classification algorithms, such as the renowned ImageNet dataset with 1.2 million images. We used a number of smaller datasets, each with considerably less than one million photos, to evaluate the applicability of CNNs in more realistic and compact contexts.

In addition, our research investigates transfer learning, looking into ways to convert models developed on sizable camera trap datasets to smaller ones. Although transfer learning has been used in previous studies (Gomez Villa et al., 2017; Norouzzadeh et al., 2018), our approach is different in that we transfer knowledge from models trained for an identical goal (animal identification) rather than from datasets that weren't collected using cameras, like ImageNet. On citizen science sites like Zooniverse, this method may make model training more effective, especially for datasets with few labelled photos.

3.1 YOLOV5

You Only Look Once is an acronym for YOLO, which stands for the series' most recent iteration, dubbed YOLOv5 [1]. YOLOv5 stands out for its anchor-based one-stage detection mechanism and incredibly quick inference speeds [2]. Due to this breakthrough, object detection is now far more effective and useful in a variety of applications.

1. **Architecture Overview:**

Three architectures—YOLOv5s, YOLOv5m, and YOLOv5l—were chosen for our investigation. The Cross Stage Partial Network (CSPNet) serves as the framework for our strategy [3]. The YOLOv5 algorithm segments the image before down sampling and before introducing the Focus module into the backbone network. The architecture's neck is made up of a Path Aggregation Network (PAN) and a Feature Pyramid Network (FPN), which efficiently combines feature data from three different scales [40, 41]. Finally, redundant prediction bounding boxes are removed using the Non-Maximum Suppression (NMS) method.

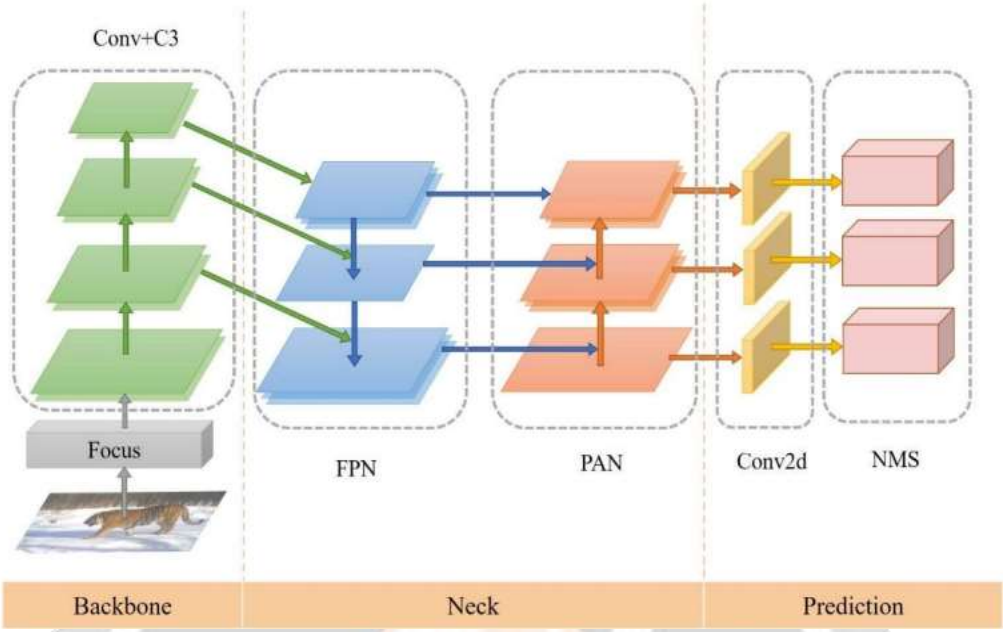

Figure 4: YOLOv5 structure diagram.

2. Implementation Details:

To train the model, we used the YOLOv5 framework and PyTorch's capabilities [42]. We used stochastic gradient descent (SGD) as our optimisation method, configuring the momentum parameter to be 0.937 and the weight decay to be 0.0005. The initial learning rate was set to 1 102, and it dropped linearly after that. We used a three-epoch-long warm-up phase with an initial warm-up momentum of 0.8 throughout training. It is important to note that the overall number of epochs and batch sizes varied due to variations in model sizes. Please see Table 1 for the precise settings of each model. Our tests were carried out on the RTX A4000 GPU.

Table 1. YOLOv5 parameter settings.

Model	Epoch	Batch Size
YOLOv5s_day	80	32
YOLOv5m_day	80	32
YOLOv5l_day	80	16
YOLOv5s_night	65	32
YOLOv5l_night	65	32

3.2. EVALUATION METRICS:

As the main evaluation criteria in this study, we used precision, recall, and mean average precision (mAP):

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

- Counting accurate detections of the ground-truth bounding box, or the number of intersections over unions (IoU) that exceed the threshold and are correctly classified, is known as "True Positive" (TP).
- False Positive (FP) refers to the number of inaccurate detections, which may comprise picking up on an object that isn't there or picking up on an object that is already there in the wrong place. This refers to the quantity that does not surpass the threshold or the quantity of incorrect classifications.
- False Negative (FN), which represents the number of missed detections and unpredicted bounding boxes, is the opposite of a positive result.

Our preferred evaluation metric for video detection scenarios was accuracy. We used a majority vote process based on the most common detection results across all frames in the target video to determine the final label for each video clip. Only when these detections had confidence levels higher than the required score threshold were they taken into account.

$$Accuracy = \frac{N}{P}$$

In this case, "N" stands for the count of correctly classified movies, while "T" stands for the overall number of videos in the dataset.

4. RESULTS

4.1. NTLNP Dataset:

We carefully examined and cleaned the data before collecting the NTLNP collection, which had 25,657 photos covering 17 different species categories. This dataset, which consists of 10,344 nighttime photos and 15,313 daytime images, was painstakingly assembled. According to Table 2, the photos in the dataset had a resolution of either 1280 x 720 or 1600 x 1200 pixels. As shown in Table 3, the NTLNP dataset was split into a training set and a test set after an 8:2 ratio split, with each set having various categories of data.

Table 2 . The main properties of the NTLNP dataset

Species Category	No. of Total Images	No. of Daytime Images	No. of Nighttime Images	Image Resolution
17	25,657	15,313	10,344	1280 × 720/1600 × 1200

Table 3. NTLNP dataset and per-class training set and test set assignments.

Species	Day and Night		Day		Night	
	Training Set	Test Set	Training Set	Test Set	Training Set	Test Set
Amur tiger	1123	246	676	145	447	101
Amur leopard	1260	314	872	219	388	95
Wild boar	1801	423	1159	291	642	132
Sika dear	1726	466	1216	328	510	138
Red fox	1504	358	802	188	702	170
Raccoon dog	1169	324	248	81	921	243
Asian badger	1052	257	735	176	317	81

Asian black bear	1084	285	772	188	312	97
Cow	1016	284	936	263	80	21
Dog	1150	280	1056	252	94	28
Total	12885	3237	8472	2131	4413	1106

4.2. Species Detection and Classification:

We chose three models that performed particularly well for the thorough assessment of species recognition accuracy: YOLOv5m, FCOS_Resnet101, and Cascade_R-CNN_HRNet32. Notably, these photographs were excluded from the model evaluation due to the data's restricted availability, with only 20 daytime images of hares being available.

The following recognition accuracies were noted in the context of species recognition for the 16 remaining species based on daylight datasets:

- The accuracy range of Cascade_R-CNN_HRNet32 was outstanding, ranging from 91.6% to 100%.
- YOLOv5m displayed accuracy between 94.2% and 99.5%.
- The accuracy range for FCOS_Resnet101 was 94% to 100%.

For the Amur leopard and musk deer, Cascade_R-CNN_HRNet32 achieved a surprising 100% recognition accuracy, while FCOS_Resnet101 excelled with 100% accuracy for the Amur tiger and red fox. In particular, YOLOv5m and FCOS_Resnet101 outperformed Cascade_R-CNN_HRNet32 by 4.4% to 4.8% when it came to the raccoon dog species, achieving recognition accuracies of 96% and 96.4%, respectively. Although YOLOv5m achieved the relatively highest accuracy of 94.2%, Sable demonstrated the lowest performance.

All of the models showed that they were capable of accurately detecting every object in a single image. It's crucial to remember that in the dataset, occurrences of multiple species showing up in front of a single camera trap simultaneously were quite uncommon. As a result, the photos in our dataset frequently featured either a single object or several objects of the same species.

Figure 5 displays a selection of the detected photos for visual reference. The Supplementary Materials section also contains additional findings made utilising the various models.

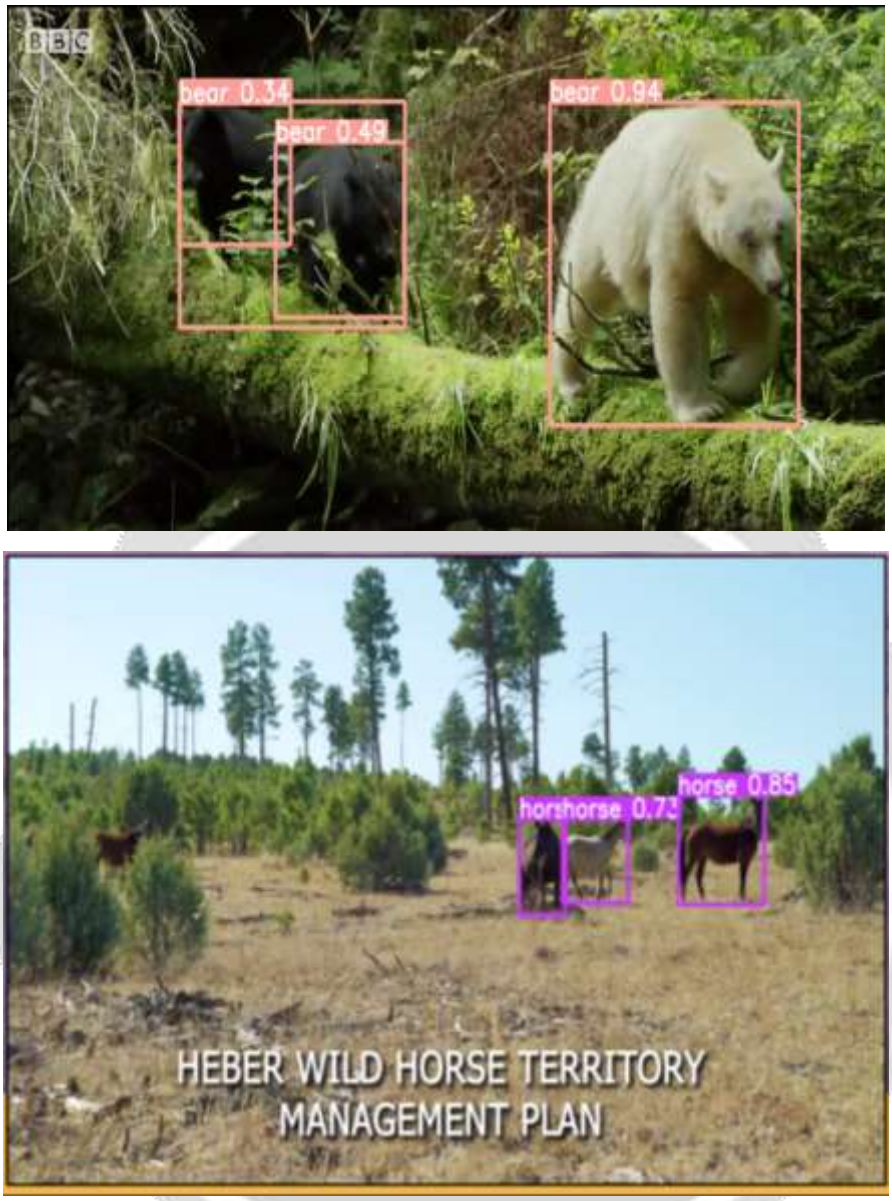

Figure 5. Examples of correct detection and classification

4.2.1. Video Automatic Recognition:

Using infrared cameras in the Northeast Tiger and Leopard National Park, we experimented with the day-night joint YOLOv5m, Cascade_R-CNN_HRNet32, and FCOS_Resnet101 models to detect objects automatically. These three models' accuracy was assessed at confidence score thresholds of 0.6, 0.7, and 0.8. Table 4 provides a summary of the findings.

YOLOv5m demonstrated the most reliable and consistent performance of the evaluated devices. It attained an accuracy of 89.6% at a confidence score threshold of 0.7. Comparatively, Cascade_R-CNN_HRNet32 fared marginally worse, reaching its maximum accuracy of 86.5% at a threshold of 0.8.

However, there were significant variances in FCOS_Resnet101's accuracy across various confidence score criteria. It achieved a video classification accuracy of 91.6% at a threshold of 0.6. However, when the threshold was raised to 0.8, the recognition rate of the videos fell precipitously and eventually only reached 64.7%.

Table 4. Video classification accuracy of the three models

Videos	Model	Acc_0.6	Acc_0.7	Acc_0.8
725	YOLOv5m	88.8%	89.6%	89.5%
	Cascade_R-CNN_HRNet32	86.3%	86.4%	86.5%
	FCOS_Resnet101	91.6%	86.6%	64.7%

5. CONCLUSION:

Studies have examined how noisy labels affect the categorization of animals. We have created a novel method for building a precise animal species categorization network using these examples of noisy labels. We looked into the network training procedure using clean samples and without them. These experiments' findings show that our noise-labeling method is accurate both with and without clean samples.

Following post-training and testing using custom datasets, the customized model yielded promising results. The overall accuracy achieved with the custom datasets was 82%. A large percentage of relevant occurrences could be properly identified by the model, as seen by the recall score, which rose to an amazing 81%. The model performed well overall, as evidenced by the F1-score's calculation of 73%, which strikes a compromise between precision and recall. It is significant to note that although the precision score was slightly lower at 66%, this can be attributed to the custom nature of the model and the constrained quantity of the training datasets.

This study emphasizes the value of including network diversity to produce a more accurate overall evaluation of sample label performance. We used k-means clustering along with deep neural network features to produce groups with a variety of traits. Groupings were then created using these clusters. Each group was then used to train its own network, making sure that every network was trained using a different collection of images. We used a maximum voting strategy to identify the real label of the noisy data.

For comprehensive wildlife monitoring carried out by citizen scientists, the suggested method for classifying animal species from camera trap photographs with noisy labels may prove invaluable (Fegraus et al., 2019). Inaccuracies in their annotations are predicted given that the majority of camera-trap photographs are gathered, examined, and shared by amateur volunteers or citizen scientists. We can extract useful animal species classifiers from these datasets using the methods we recommend.

Supplementary Materials: The experiment's source code can be found at: https://github.com/saravanan-2003/EcoScan-AI-powered-Animal-Recognition-and-Species-Categorization (accessed on 01 June 2023).

Data Availability Statement: NTLNP_dataset link: https://pan.bnu.edu.cn/l/s1JHuO (accessed on 1 May 2023).

References:

[1] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2019; pp. 779–788.

[2] Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. arXiv 2020, arXiv:2004.10934.

[3] Wang, C.-Y.; Liao, H.-Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–7 June 2020; pp. 1571–1580

[4] Padilla, R.; Netto, S.L.; Da Silva, E.A. A survey on performance metrics for object-detection algorithms. In Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niterói, Brazil, 1–3 July 2020; pp. 237–242.

[5] L.L Ricky, E.M William, Deaths resulting from animal attacks in the United States, Wilderness Environ. Med. 8 (1) (1997) 8–16 doi:10.1580/1080-6032(1997)008[0008:DRFAAI]2.3.CO2.

[6] R.M. Nowak, in: Walker's Mammals Of The World, 1, 6th Edition, Johns Hopkins University Press, Baltimore, 1999, pp. 1166–1170

[7] L Karlinsky, S Joseph, H Sivan, S Eli, A Amit, F Rogerio, G Raja, M.B Alex, RepMet: representative-based metric learning for classification and few shot object detection, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019, June 2019, doi:10.1109/cvpr.2019.00534

[8] Chen, G.; Han, T.X.; He, Z.; Kays, R.; Forrester, T. Deep convolutional neural network based species recognition for wild animal monitoring. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 858–862.

[9] 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. Proc. Natl. Acad. Sci. USA 2018, 115, E5716–E5725.

[10] Schneider, S.; Taylor, G.W.; Kremer, S. Deep learning object detection methods for ecological camera trap data. In Proceedings of the 2018 15th Conference on Computer and Robot Vision (CRV), Toronto, ON, Canada, 8–10 May 2018; pp. 321–328.

[11] Zhao, Z.-Q.; Zheng, P.; Xu, S.-t.; Wu, X. Object detection with deep learning: A review. IEEE Trans. Neural Netw. Learn. Syst. 2019, 30, 3212–3232.

[12] Burton, A.C.; Neilson, E.; Moreira, D.; Ladle, A.; Steenweg, R.; Fisher, J.T.; Bayne, E.; Boutin, S. Wildlife camera trapping: A review and recommendations for linking surveys to ecological processes. J. Appl. Ecol. 2015, 52, 675–685.

[13] Yu, X.; Wang, J.; Kays, R.; Jansen, P.A.; Wang, T.; Huang, T. Automated identification of animal species in camera trap images. EURASIP J. Image Video Process. 2013, 2013, 52

[14] Sahu, R. Detecting and Counting Small Animal Species Using Drone Imagery by Applying Deep Learning. In Visual Object Tracking with Deep Neural Networks; IntechOpen: London, UK, 2019.

[15] Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. Insights Imaging 2018, 9, 611–629.

[16] Brownlee, J. How to Configure Image Data Augmentation in Keras.

Available online: https://machinelearningmastery.com/ how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/ (accessed on 8 June 2023).

[17] Wu, R.; Yan, S.; Shan, Y.; Dang, Q.; Sun, G. Deep image: Scaling up image recognition. arXiv 2015, arXiv:1501.02876.

[18] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, Adv. Neural Inf. Process. Syst. (2014) 487–495

[19] Z. Zhang, Z. He, G. Cao, W. Cao, Animal detection from highly cluttered natural scenes using spatiotemporal object region proposals and patch verification, IEEE Trans. Multimed. 18 (10) (2016) 2079–2092.

[20] B. Yuan, J. Chen, W. Zhang, H.S. Tai, S. McMains, Iterative cross learning on noisy labels, in: In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, March, pp. 757–765.