ECO AI'S AUTOMATED ANIMAL IDENTIFICATION AND DETECTION PROJECT

HARIHARAN B[1], KARTHIKEYAN N[2], BALAKRISHNAN K[3], ESAKKI MADURA E[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] *Assistant Professor -I, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India*

ABSTRACT:

Detecting and classifying animal species serves as a foundational step in assessing their long-term viability and the impact our actions may have on them. Additionally, this process aids in the recognition of both predatory and non-predatory animals, both of which pose substantial threats to both humans and the environment. Moreover, it contributes to the reduction of traffic accidents in various regions, where animal encounters on roadways have led to numerous automobile collisions. Nonetheless, the task of detecting and classifying animal species is fraught with challenges, including variations in size and disparate behaviors among species. This paper presents an innovative approach, employing a novel two-stage network with a modified multi-scale attention mechanism, to create an integrated system that effectively addresses these challenges. At the regional proposal stage, we adopt a pyramid design with lateral connections, enhancing the sensitivity of semantic characteristics for smaller objects. Furthermore, we employ a densely connected convolutional network to enhance functional transmission and multiplex it throughout the classification stage, resulting in more precise classification with fewer parameters. Our project demonstrates that deep neural networks, a cutting-edge form of artificial intelligence, can autonomously extract such data. The ultimate goal is to train neural networks for automatic animal identification and recognition, a step forward in harnessing the potential of these technologies.

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

1. INTRODUCTION:

Detecting and classifying animal species plays a crucial role in addressing various challenges, including wildlife-related road accidents resulting in fatalities and injuries and human-wildlife conflicts (Nowak et al., 5). Animal attacks, responsible for numerous human injuries and fatalities, exhibit varying frequencies depending on geographic regions. For instance, in the United States, an estimated two million animal attacks on humans occur annually (Warrell, 6). Tanzanian and American scientists report a notable increase in such incidents from 1990 to 2005, with at least 563 villagers falling victim to animal attacks during this period. Predatory animals, such as tigers and lions, are known to pose significant risks, with tigers accounting for a higher number of human fatalities than any other species of their kind (Nowak et al., 5). However, the lack of comprehensive records across governments obscures the true extent of animal-related deaths. Animal attacks often occur during the night due to hunger, as animals venture in search of food, making the development of effective techniques for animal detection, classification, and monitoring crucial to mitigate these risks, prevent animal-vehicle accidents, and deter theft. Object detection, a rapidly evolving field in computer vision, is central to these efforts, with deep learning techniques like CNNs showcasing exceptional performance in image comprehension. Two-stage detectors, such as Faster R-CNN, R-FCN, FPN, and YOLOv5 (Birds class, 7), have gained significant attention due to their high precision. Challenges related to anchor sizes persist, impacting the accuracy of detection. In the realm of computer vision, the application of animal detection is crucial for solving diverse challenges, including wildlife accidents and the protection of endangered species (Birds class, 7). Recognizing animals poses unique challenges, primarily related to variations in shape, color, and appearance within the same species (Birds class, 7). Differences in lighting conditions and orientations also affect animals' identification. These challenges require specialized models with significant learning capabilities to identify numerous animal breeds in still images, utilizing convolutional neural networks (CNNs) with fewer parameters and connections (Birds class, 7). Attention mechanisms in object detection and classification frameworks, including intricate and soft attention, have also garnered attention, along with the Region Proposal Network (RPN) to handle tiny animal species. While approaches involving image enlargement and high-resolution detection maps are used to address small animal detection, multi-level representation network variations enhance model capability. The pursuit of real-time applications, however, presents computational challenges in addressing these issues effectively (Birds class, 7).

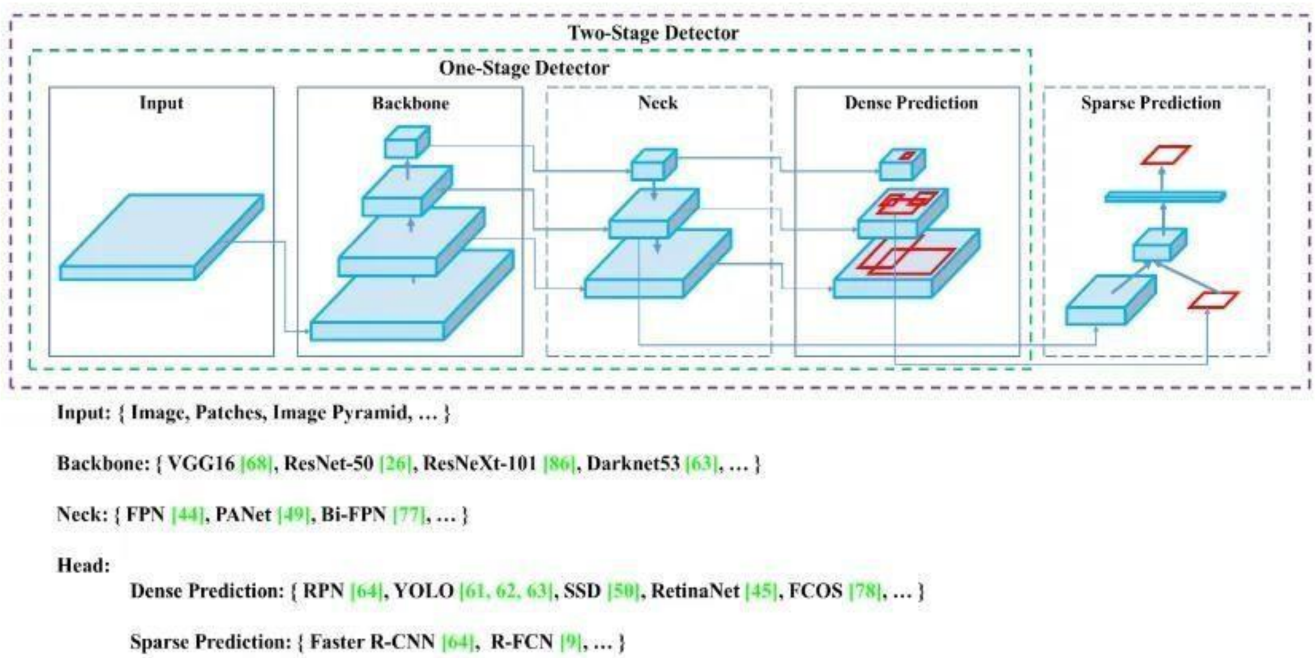

Figure 1 Object Detection

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:

In the realm of deep learning, particularly within the domain of supervised learning, the fundamental objective revolves around mastering the art of mapping input data to desired output categories through the application of specialized neural network architectures (Goodfellow et al., 2016). In the context of image classification, the primary goal is to train a deep learning algorithm capable of processing and categorizing images into predefined classes, including distinct animal species. Over the recent years, the field of image classification has witnessed the profound rise of convolutional neural networks

(CNNs), with their dominance exemplified in notable challenges like the ImageNet Large Scale Visual Recognition Challenges (ILSVRC) (Krizhevsky, Sutskever, & Hinton, 2012; Russakovsky et al., 2015).

CNNs, initially introduced by LeCun et al. in 1989, comprise two interconnected core components: a convolutional section designed to extract localized features from images and a fully connected segment responsible for mapping these acquired features to the desired output categories (LeCun et al., 1989). Unlike earlier approaches, CNNs eliminate the necessity for manually crafted features. Instead, they autonomously acquire spatial features by adjusting their parameters (weights) during model training, accomplished by propagating errors from the output layer back to the input. The precise configuration of operations applied to the data within a CNN defines its architecture. Figure 3 schematically illustrates the architecture of a CNN, highlighting its primary unit, a layer, which includes filters conducting convolutions on the input data to discern spatial patterns, incorporate activation functions, and perform pooling (sub-sampling) operations. Each layer typically generates smaller feature maps, subsequently forwarded to the subsequent layer. Multiple such layers are typically arranged sequentially, enabling intricate feature extraction. The count of layers within a neural network's architecture defines its depth, signifying the essence of deep learning – neural networks endowed with numerous layers (He et al., 2015).

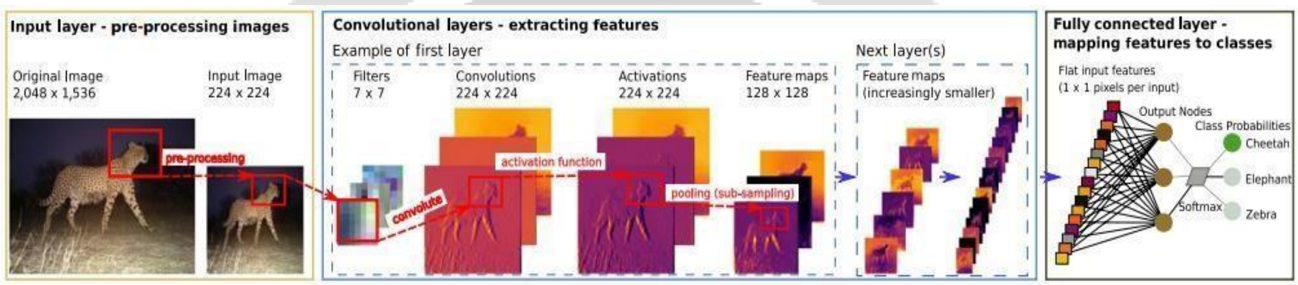

Figure 3 Schematic illustration of CNN architecture

2. METHODOLOGY

Our classifier operates in two distinct phases: training and testing. During the training phase, a collection of images serves as visual exemplars. In the subsequent testing phase, a freshly captured image, referred to as the test image, is presented as input to the classifier. Leveraging the insights acquired during training, the classifier then categorizes the test image into the most appropriate class.

A. *Receiving the input image:*

Within the envisioned system, an image is acquired through the camera connected to the system. This captured test image serves as the initial input and undergoes conversion into a binary pattern. Subsequently, a dataset containing previously labeled images is utilized, and their distinctive features are compared with those extracted from the test image. This comparison process aids in identifying the specific animal species within the image.

B. *Feature Extraction*:

The input test image can be processed to generate a condensed set of features. These chosen features may encompass crucial information from the input data, enabling the accomplishment of the desired task with this streamlined dataset instead of the original, unmodified data. Fixed features, known as human-crafted features, are directly derived from images. In contrast, deep neural networks, unlike human-crafted features, identify features within images and establish multiple tiers of representation, with the upper-level features encapsulating more abstract aspects of the data.

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

In the context of species classification, the output layer is responsible for calculating the probabilities associated with the presence of the detected animal in the image, categorizing it into one of the potential classes. While furnishing such an outcome could significantly reduce the human effort required for accurate species identification, verifying this hypothesis still necessitates human expertise and knowledge.

3. LITERATURE SURVEY:

Initial research in automated animal identification primarily focused on matching species-specific patterns in images, requiring extensive manual preprocessing. However, the achieved accuracies, such as the 82% reported by Yu et al. (2013), fell short of human-level accuracies, which reached 96.6% (Swanson et al., 2016). Recent studies employing Convolutional Neural Networks (CNNs) for automatic animal species identification have reported accuracies around 90%, with some involving manual preprocessing (Gomez Villa et al., 2017) or more complex pipelines with automatic preprocessing (Giraldo-Zuluaga, Salazar, Gomez, & Diaz-Pulido, 2017). The most recent advancements by Norouzzadeh et al. (2018) achieved accuracies of 93.8%, matching human accuracy on over 99% of images.

Our study seeks to apply and validate CNNs across a broader range of camera trap datasets compared to previous research. While Norouzzadeh et al. (2018) demonstrated impressive results with the Snapshot Serengeti dataset, consisting of 3.2 million images, most camera trap datasets, as observed on Zooniverse, are smaller in scale. Effective image classification models often require substantial datasets, like the renowned ImageNet dataset containing 1.2 million images. To assess the applicability of CNNs in more realistic and smaller-scale scenarios, we incorporated several smaller datasets, each comprising significantly fewer than one million images.

Moreover, our work explores transfer learning, investigating how to adapt models trained on large camera trap datasets to smaller ones. While transfer learning has been applied in prior studies (Gomez Villa et al., 2017; Norouzzadeh et al., 2018), our unique approach involves transferring knowledge from models trained for a similar task (animal identification) rather than non-camera trap datasets (e.g., ImageNet). This approach holds potential for more efficient model training on citizen science platforms like Zooniverse, particularly for datasets with limited labeled images.

3.1 YOLOV5

YOLO, an abbreviation for 'You Only Look Once,' represents the cutting-edge development in the YOLO series, known as YOLOv5 [1]. Distinguished by its anchor-based one-stage detection mechanism, YOLOv5 boasts remarkably fast inference speeds [2]. This innovation has greatly enhanced the efficiency of object detection, making it a valuable asset in various applications.

1. Architecture Overview:

For our study, we selected three architectures, namely YOLOv5s, YOLOv5m, and YOLOv5l. The backbone of our approach incorporates the Cross Stage Partial Network (CSPNet) [3]. Preceding the entry into the backbone network, the YOLOv5 algorithm introduces the Focus module and conducts down sampling by segmenting the image. The neck of the architecture takes the form of a Feature
Pyramid Network (FPN) complemented by a Path Aggregation Network (PAN), effectively incorporating feature information from three distinct scales [40,41]. Finally, it employs the Non- Maximum Suppression (NMS) technique to eliminate redundant prediction bounding boxes.

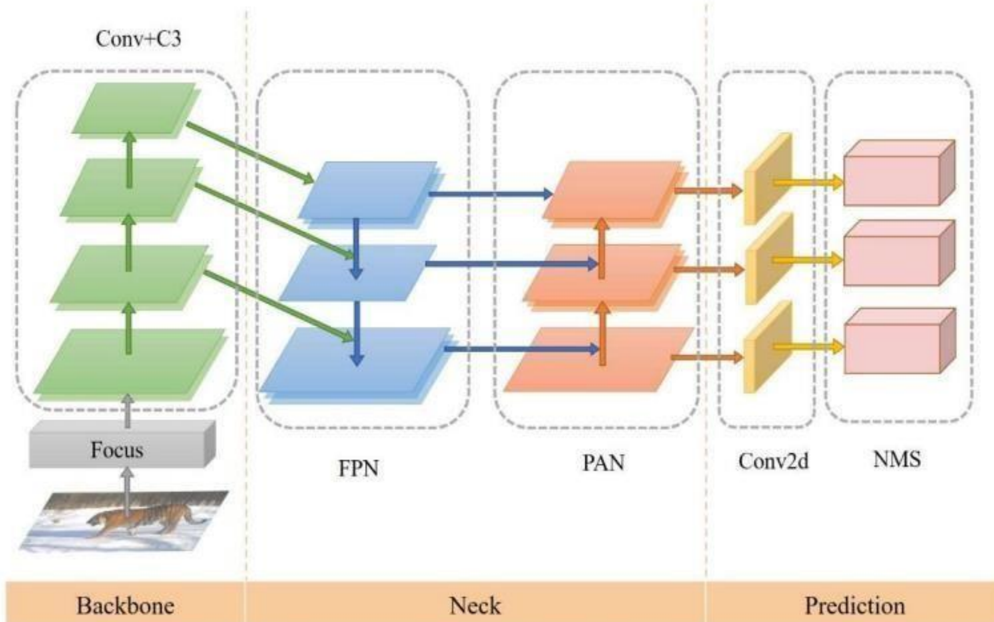

Figure 4: YOLOv5 structure diagram.

2. Implementation Details:

We employed the YOLOv5 framework for model training, leveraging the capabilities of PyTorch [42]. Our optimization strategy relied on Stochastic Gradient Descent (SGD), where the momentum parameter was configured at 0.937, and the weight decay was set to 0.0005. The initial learning rate was initialized to $1 \times 10-2$ and experienced linear decay. During training, we implemented a warm- up phase spanning three epochs, with an initial warm-up momentum of 0.8. It's worth noting that due to variations in model sizes, the total number of epochs and batch sizes differed. For specific configurations of each model, please refer to Table 1. Our experiments were conducted using 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:

In this research, we employed precision, recall, and mean average precision (mAP) as the key evaluation metrics:

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

$$TP + FP$$

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

- True Positive (TP) represents the count of accurate detections of the ground-truth bounding box, signifying the number of intersections over union (IoU) that surpass the threshold and are correctly categorized.
- False Positive (FP) denotes the count of incorrect detections, which could involve either detecting a non-existent object or misplacing detections of an existing object. This refers to the number of that do not exceed the threshold or the number of classification errors made.
- False Negative (FN) is the count of missed detections, indicating the number of bounding boxes that were not predicted.

In video detection scenarios, our evaluation metric of choice was accuracy. To assign a final label to a video clip, we employed a majority voting mechanism based on the most frequently occurring detection results across all frames within the target video. These detections were considered only if their confidence levels exceeded the specified score threshold.

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

Here, "N" represents the count of videos that were correctly classified, and "T" represents the total number of videos in the dataset.

4. RESULTS

4.1 NTLNP Dataset:

After a thorough examination and cleaning process, we curated a dataset named NTLNP, which consisted of 25,657 images spanning 17 distinct species categories. This dataset was carefully compiled, comprising 15,313 images captured during daylight and 10,344 images taken at night. The images in the dataset exhibited a resolution of either 1280 × 720 or 1600 × 1200 pixels (as detailed in Table 2). Following an 8:2 ratio split, the NTLNP dataset was partitioned into a training set and a test set, each containing different types of data, as illustrated in Table 3.

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:

For the comprehensive evaluation of species recognition accuracy, we selected three models, namely YOLOv5m, FCOS_Resnet101, and Cascade_R-CNN_HRNet32, which exhibited superior performance. Notably, due to the limited availability of data with only 20 images of hares taken during the daytime, these images were not included in the model evaluation.

In the context of species recognition for the 16 remaining species based on daytime datasets, the following recognition accuracies were observed:

- Cascade_R-CNN_HRNet32 achieved an impressive accuracy range of 91.6% to 100%.
- YOLOv5m exhibited accuracy within the range of 94.2% to 99.5%.
- FCOS_Resnet101 demonstrated accuracy spanning from 94% to 100%.

Furthermore, Cascade_R-CNN_HRNet32 achieved a remarkable 100% recognition accuracy for Amur leopard and musk deer, while FCOS_Resnet101 excelled with 100% accuracy for Amur tiger and red fox. Specifically, in the case of the raccoon dog species, YOLOv5m and FCOS_Resnet101 achieved recognition accuracies of 96% and 96.4%, respectively, outperforming Cascade_R-CNN_HRNet32 by

4.4% to 4.8%. However, it's worth noting that Sable exhibited the lowest performance, with YOLOv5m achieving the relatively highest accuracy of 94.2%.

All of the models demonstrated the capability to successfully detect each object within a single image. It's important to note that in the dataset, instances of different species appearing simultaneously in front of a single camera trap were quite rare. Consequently, the images in our dataset typically contained either a single object or multiple objects belonging to the same species.

For visual reference, a selection of identified images is presented in Figure 5. Moreover, additional results obtained using the various models can be found in the Supplementary Materials section.

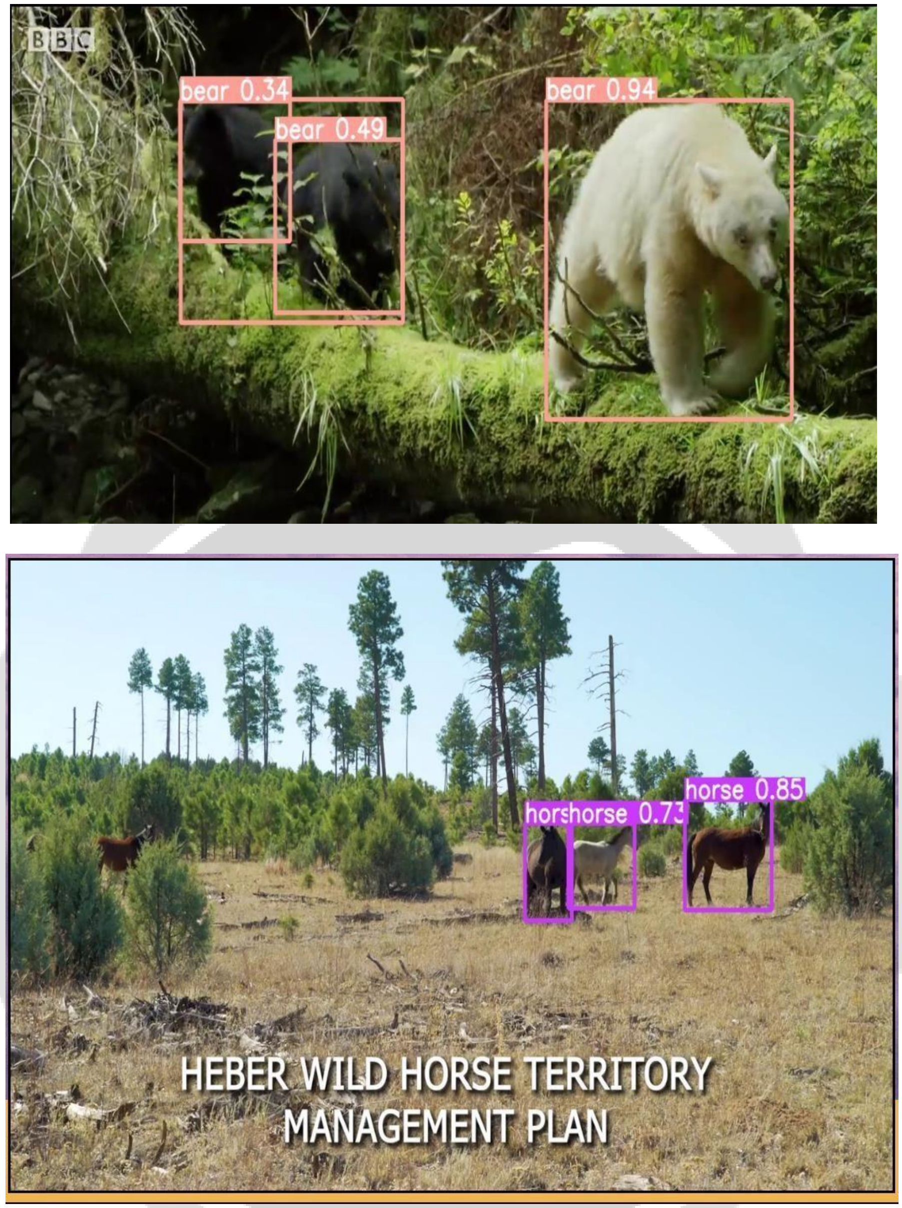

Figure 5. Examples of correct detection and classification

4.2.1. Video Automatic Recognition:

We conducted experiments using the day-night joint YOLOv5m, Cascade_R-CNN_HRNet32, and FCOS_Resnet101 models to automatically recognize videos captured by infrared cameras within the Northeast Tiger and Leopard National Park. The accuracy of these three models was evaluated at different confidence score thresholds: 0.6, 0.7, and 0.8. The results are summarized in Table 4.

Among the models tested, YOLOv5m exhibited the most consistent and robust performance. At a confidence score threshold of 0.7, it achieved an accuracy of 89.6%. In comparison, Cascade_R- CNN_HRNet32 performed slightly less effectively, achieving its highest accuracy of 86.5% at a threshold of 0.8.

However, the accuracy of FCOS_Resnet101 showed notable variations across different confidence score thresholds. At a threshold of 0.6, it achieved a video classification accuracy of 91.6%. Nevertheless, as the threshold was increased to 0.8, the recognition rate of the videos experienced a sharp decline, ultimately reaching only 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 delved into the impact of noisy labels on animal classification. From these noisy label examples, we have developed an innovative technique for constructing a precise animal species categorization network. We investigated the network training process with and without clean samples. The results of these studies highlight the accuracy of our noise-labeling method, 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%. Additionally, the recall score reached an impressive 81%, indicating the model's ability to correctly identify a high proportion of relevant instances. The F1-score, which balances precision and recall, was calculated at 73%, demonstrating the model's balanced performance. It's important to note that while the precision score was somewhat lower at 66%, this can be attributed to the model's custom nature and the limited size of the datasets used for training.

This research underscores the significance of incorporating network diversity to achieve a more precise collective assessment of sample label performance. To create groups with diverse characteristics, we harnessed deep neural network features coupled with k-means clustering. These clusters were then used to form groupings. Subsequently, each group was employed to train its own network, ensuring that each network received training using a unique set of images. To determine the true label of the noisy data, we applied a maximum voting approach.

The proposed method for categorizing animal species from camera trap photos with noisy labels could prove invaluable for extensive wildlife monitoring conducted by citizen scientists (Fegraus et al., 2019). Given that most camera-trap photos are collected, analyzed, and shared by amateur volunteers or citizen scientists, inaccuracies in their annotations are expected. Using our suggested methodology, we can extract effective animal species classifiers from these datasets.

Supplementary Materials: The source code of the experiment is available 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.