Animal Species Detection and Classification Framework Based on Modified Multi-Scale Attention Mechanism and Feature Pyramid Network

Article *in* Scientific African · March 2022

DOI: 10.1016/j.sciaf.2022.e01151

CITATIONS
15

READS
553

9 authors, including:

Chiagoziem Chima Ukwuoma
University of Electronic Science and Technology of China
61 PUBLICATIONS **556** CITATIONS
SEE PROFILE

Sophyani Banaamwini Yussif
University of Electronic Science and Technology of China
21 PUBLICATIONS **79** CITATIONS
SEE PROFILE

Happy Nkanta Monday
Oxford Brookes College of Chengdu University of Technology
55 PUBLICATIONS **596** CITATIONS
SEE PROFILE

Grace U. Nneji
Oxford Brookes College of Chengdu University of Technology
49 PUBLICATIONS **447** CITATIONS
SEE PROFILE

Scientific African 16 (2022) e01151

Contents lists available at ScienceDirect

Scientific African

journal homepage: www.elsevier.com/locate/sciaf

Animal species detection and classification framework based on modified multi-scale attention mechanism and feature pyramid network

Chiagoziem C. Ukwuoma[a,*], Zhiguang Qin[a,*], Sophyani B. Yussif[b], Monday N. Happy[b], Grace U. Nneji[a], Gilbert C. Urama[b], Chibueze D. Ukwuoma[c], Nimo B. Darkwa[b], Harriet Agobah[a]

[a] School of Information and Software Engineering, University of Electronic Science and Technology of China, Sichuan PR China
[b] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Sichuan PR China
[c] Departmemt of Physics-Electronics, Federal University of Technology Owerri, Nigeria.

ARTICLE INFO

Article history:
Received 20 August 2021
Revised 17 January 2022
Accepted 15 March 2022

Editor: DR B Gyampoh

Keywords:
Deep Learning
Multiscale Attention Mechanism
Feature pyramid
Animal Detection
and Classification

ABSTRACT

Detecting and classifying animal species is the first step in determining their long-term viability and the influence we may be having on them. Second, it aids people in recognizing predators and non-predatory animals, both of which pose a significant threat to humans and the environment. Third, it lowers the rate of traffic accidents in various regions since it has been a regular sighting on roadways, resulting in several collisions with automobiles. However, animal species' detection and Classification of animal species face many challenges such as the size and inconsistent behaviors various among the species. This paper proposes using a novel two-stage network with a modified multi-scale attention mechanism to create a more integrated recognition and classification system to attend to the challenges. At the regional proposal stage, a deeply characterized pyramid design with lateral connections was adopted, making the semantic characteristic of a small item more sensitive. Secondly, by reason of a densely connected convolutional network, the functional transmission is enhanced and multiplexed throughout the classification stage, resulting in a more precise Classification with fewer parameters. The Proposed model was evaluated using the AP and mAP evaluation metrics on the Animal wildlife and the challenging Animal-80 dataset. An mAP of +0.1% and an AP of 5% to 20% increase in each class was achieved by the attention-based proposed model compared to the non-attention-based model. Further comparison with other related works shows the proposed techniques' effectiveness for detecting and classifying animal species.

* Corresponding author.
E-mail addresses: ukwuoma@std.uestc.edu.cn (C.C. Ukwuoma), qinzg@uestc.edu.cn (Z. Qin).

https://doi.org/10.1016/j.sciaf.2022.e01151

Introduction

Detection and Classification of animal species is an area that needs good techniques as it reduces the problems of wildlife road accidents leading to deaths and injuries and helps humans understand diversity better. Animal assaults frequently cause most human fatalities and injuries. The frequency of animal assaults varies depending on where you live. For instance, in the United States, an estimate of two million animals' assaults on humans are recorded each year [38]. In contrast, Tanzanian and American scientists' reports show that Animal assaults on humans increase from 1990 to 2005. The report stated that at least 563 villagers were being attacked and eaten during this time. Predator animal is a colloquial word describing a species of animal that hunts humans as prey. Tigers, for example, are known to murder more humans than any other animal of their type, as also recorded by Nowak et al [56].. Lions are not left out as report also shows that they attack people living in rural areas at night as well as daytime in search of food. Warrell [57] claims that animal assaults cause tens of thousands of deaths each year throughout the world. However, it does not appear that every government keeps records of animal-related deaths. Most animals do not attack humans daily, except for tigers, while certain animals feed on sick, unconscious, or dead people. When animals become used to humans or are severely starved, they may attack pets, livestock, and people. Animal attacks are most common during the night as a result of hunger. Thus the animals roam about in search of food. Excellent and accurate techniques are needed to detect, classify, and monitor animals more efficiently, thus preventing animal-vehicle accidents, tracing animals, and preventing theft. One fast-growing and typical computer vision task is Object Detection. Deep learning techniques like CNNs have lately been shown to accomplish various image comprehension by the outstanding performance of recent research results. Existing object detectors could also be grouped into two types, one-stage and two-stage detectors. Different anchor sizes are employed to predict the target bounding boxes for one-stage detectors, such as the RetinaNet [1], SSD [2], YOLO9000 [3], and YOLOv2 [4], which is similar to the beginning phase of the Faster CNN [5]. Despite the excellent performance in terms of speed of the one-stage detectors, there exists a research gap; when we have a smaller anchor size, the target bounding box is missing, however when the anchor size is too large, the wider receptive field reduces the actual properties of the target resulting to a shallow score causing poor performance. With the two-stages detector, like Faster R-CNN [5], R-FCN [6], and FPN [7], the regional proposal networks (RPN) and the classification networks are all involved. The two-stage framework is the general focus of researchers in present-day object detectors because of its incredible precision.

Computer Vision (CV) approaches for animal detection are essential, adding to other animal recognition and classification approaches in solving unique problems such as wildlife accidents and endangered species [8,9]. CV has become a widely employed tool in imagery, health, vehicle mapping, and drones in modern culture [10,11,49]. These applications are set up to realize objects, such as localization, detection, and Classification [12,13,51]. Significant differences in shape and color appearance from different objects of the same class are essential issues affecting object detectors' efficiency in image processing [14,15,52]. Unlike the body and face, the distinct positions of the body and shape of the animal body have more differences in appearance that are almost normal and unique [16,17]. Contrasting light intensities and orientations also affect animals' identification [18]. The Birds class [19] states that organisms of a similar group may considerably differ in other properties. The recognition of several animals potentially means that one animal has to differentiate between others of the same kind. These issues had brought about several human problems before technology was employed [20]. Moreover, the identification of animal heads has obvious obstacles, as the face of animals varies significantly from the face of humans. A particular creature may swiftly take up different forms and colors to make it challenging to identify. The current development of neural networks has significantly upgraded these latest visual recognition systems [21,22]. A unique model with tremendous learning ability is needed to learn about thousands of animal breeds in still images. Therefore, this model must have masses of primary data to offset all the missing data sets [23]. This results in the CNN having considerably fewer parameters and connections than standard feedforward neural networks with equivalent layers. Their training and testing will be more straightforward, but theoretically, their optimal performance is likely to be slightly inferior [24,25]. Despite the trials to develop a more robust architecture with the ability to recognize and classify animal species [26–29], some studies have focused on the use of attention mechanism on object detection and classification frameworks, which often may be grouped into intricate- attention [28,30] and soft- attention [27]. The Region proposal Network (RPN) achieved state-of-the-art performance in object detection. It focuses on only the valuable part of the image and discards the part without the needed information. Attempts were made [31,2,33-35] as of late in two approaches to deal with the issue of tiny object detection (Tiny animal species). First, small animal species will be handled discretely by enlarging the image size and then making high-resolution detection maps, which will require additional training and testing time, thus straining the detection apps in real-time. A different way of creating multi-level representation network variations can improve the model's ability [53–55], but the computational load for real-time applications is still a problem.

This paper proposed a two-stage network (detection stage and classification stage) to deal with animal species detection and classification challenges, as illustrated in Figure 1. The ResNet-50 network serves as our model backbone network for the detection stage. We built an attention mechanism and a feature pyramid network to enhance the network's responsiveness to tiny targets without significantly expanding the networks' complexity. The feature pyramid increases the characteristic responses by plainly modeling interdependencies between the channel characteristics to capture channel-wise relationships and give discriminatory object recognition functions. The regression of bounding boxes and the Classification of tiny species have enhanced yields by picking more discriminatory features based on the attention mechanism. We used the Focal Loss as the loss function to further improve the network performance to supervise the region proposal network. For the second

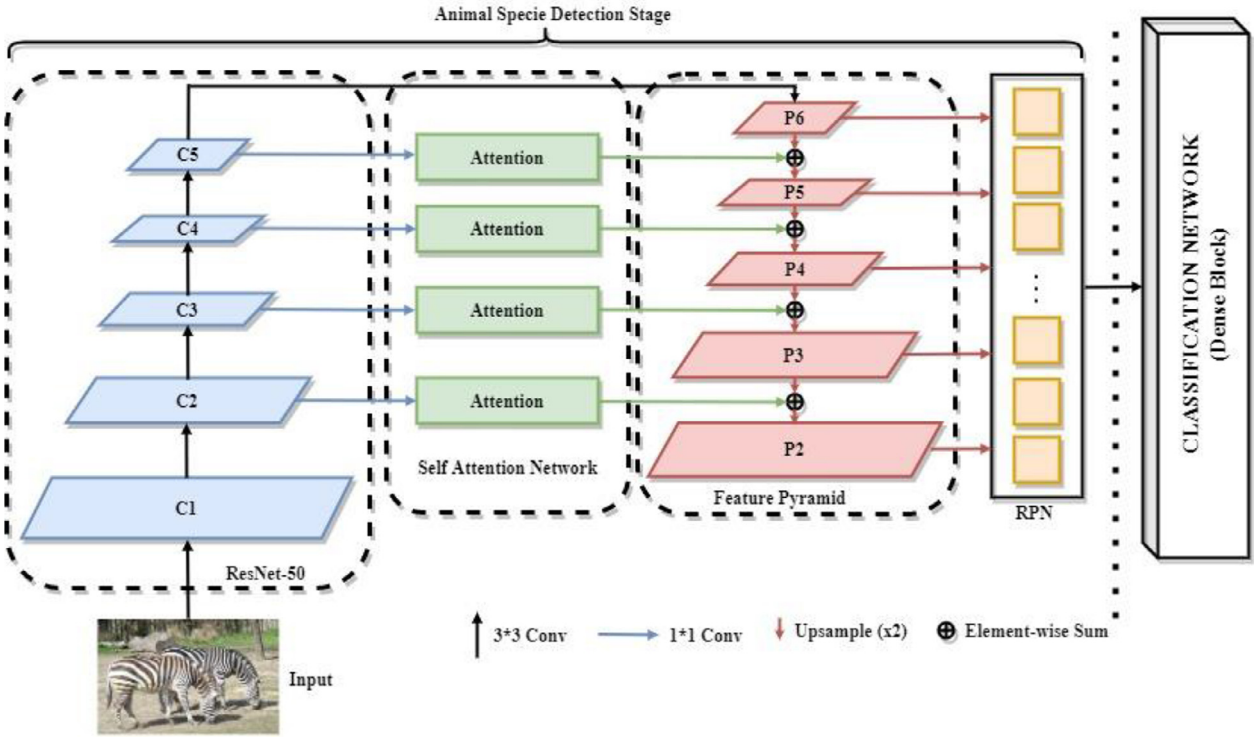


Figure 1. Proposed Animal Species Detection and classification Network

stage, the classification stage, dense blocks boost network classification accuracy. It is worthy to note that we turned our input into a two-tiered image pyramid which is helpful for big animal species detection. The significant contribution of this paper is summarized as;

- v This paper proposed a novel, simple, yet effective framework for animal species detection and Classification.
- v This paper proposed using ResNet-50 Network as a backbone to design a multiscale attention mechanism that does dot-multiplication and SoftMax by multiscale features to receive the weights, thus summing it to finetune the features of the species for better detection and Classification.
- v We further designed a future pyramid to extract the finetune attention features of the animal species and modify the anchors to fit into the small species for better performance.
- v The focal Loss is used as the loss function and dense block in the classification network to boost the accuracy of our proposed network.
- v Lastly, a detailed experiment was done on two challenging Animal datasets using the AP and mAP evaluation metrics. The results support the argument that the proposed model solves the earlier mentioned problems of Animal species detection and Classification.

The remainder of this paper is summarized as follows;

We discussed the related works covering Animal species detection and Classification, generic object detection, and attention mechanism in image processing in Section 2. In contrast, Section 3 discusses the materials and methods such as the proposed model, pyramidal feature network, attention mechanism, architecture details, experiments, implementation details, data preprocessing, and evaluation metrics. We present our findings, limitations of our proposed method, and future works in section 4, and finally, we conclude in Section 5.

Related work

This section presents the related works on animal species detection and Classification, generic object detection, and attention mechanism in image processing.

Animal species detection and classification

The usual machine-learning tools used in species-distribution modeling consist of decision trees, widespread linear models, fluctuating logic, genetic algorithms for generating rules set, maximum entropy methods, and the random forest. Chosen

and manually extracted images with an entire animal shape always yield a high accuracy result. The discoveries of the conventional machine learning algorithms such as SIFT combination, local-structured binary patterns, sparse weighted dictionary education coding, and Support Vector Machine (SVM) were published in [36]. However, this conventional algorithm requires some special features if the animal is previously known. Still, the animal images are also manually cropped to select only the whole animal shape before feeding to the conventional machine learning algorithm. They are mainly applied to individual identification and Classification.

The birth of deep learning has greatly affected the visual detection and classification systems seen in [37] as a state-of-the-art approach for recognizing wildlife. Villa et al [39]. worked on animal species identifying structures utilizing a digital camera connected to an infrared sensor to capture animal images using the AlexNet, VGGNet, GoogLeNet (Inception), and ResNets. The Blackbox extractor technique of ConvNets was then applied. Wildlife spotter, Australia, was explored by [37]. The Wildlife Detector was given as a CNN model that trains a binary classification (with two classes, animal and non-animal) and wildlife identification as another CNN model, which would learn a multi-class classifier (species identification). The "so-called" Lite Alex Net was a wildlife detector, and two ConvNets (VGG-16 and ResNet-50) had lesser hidden layers and feature maps on each layer. The camera-trap images form a mix of favorite ways to segment and classify animals at times [40,58]. has employed animal recognition techniques in the Colombian forest, by the name of robust layer principal component analysis for segmentation, CNN for extracting features, the LASSO (Least Absolute Shrinkage and Selection Operator) for characteristics, and the SVM for Classification of mammalian genera. The CNN was applied as GoogLeNet, ResNet50, ResNet101, ResNet152, and MixtureNet. The authors obtained significantly excellent accuracy values contributing to CNN, LASSO, and SVM applications.

Generic object detection

Ross et al [41]. proposed an RCNN network employing selective search to obtain the region proposal. It merges region proposals with CNN to find the object. The residual network [42] with units adds an Identity mapping to aid in alleviating the deterioration of the model. It was utilized for different broad domains and is commonly used as a backbone network for generic object identification. Faster R-CNN based on the RPN was proposed by [5] to create a region proposal, extract features, and categorize and exact object locations. It dramatically increases speed and precision. Lin et al [7]. proposed the multi-scale object detecting feature pyramid network. The intrinsic multi-stage pyramidal levels of CNN and top-down network topology with oblique connections produce feature pyramids with additional marginal costs. Each high-level map with semantics data is developed as a substantially improved generic feature extractor. By depending on Faster RCNN, Mask R-CNN [45] extends by a branch to parallel the prediction of a masked object with an existing unit. Mask RCNN gives simultaneously high-quality semantic segmentation and efficient object identification. Cascade R-CNN [46] utilizes numerous IoU thresholds to train several waterfall detectors for noise interference and inaccurate object identification issues.

Attention in image processing

Techniques based on attention mechanisms emphasize crucial information and disregard inconsistency. Mnih et al [47]. recommended a recurring RNN-based attention model for the first time. It adaptively picks the series of areas or locations, processes just the regions, and extracts high-resolution features. It is recommended for establishing Spatial Transformer Networks (STN) [27]. The combination of the convolutional network is considered a violent approach in which direct combinations of data result in ignorance of essential information. The STN of space information in images is beneficial to get important information. The residual attention network [28] uses the channel and spatial attention, which means each piece chosen to constitute a mask is produced by weight. Ref [32]. improves the capacity to describe networks by recalibrated channel characteristics. Loss functions give the weights of the attributes. With inspiration from classical non-local methods and captures non-local operations for long-range dependence. This non-local procedure is applied in computing each function's weight and eventually summarizes all available weights. Woo et al [44]. integrated self-attention channel [43] and spatial attention [28].

Motivation

Animal species detecting and classifying is the first step in determining their long-term viability and the influence we may be having on them. Second, it aids people in recognizing predators and non-predatory animals, both of which pose a significant threat to humans and the environment. Third, it lowers the rate of traffic accidents in various regions since it has been a regular sighting on roadways, resulting in several collisions with automobiles. Research shows that traditional machine learning approaches such as SIFT combination, local-structured binary patterns, sparse weighted dictionary education coding, and Support Vector Machine (SVM) have been applied for the task of Animal species detection and classification. However, this conventional algorithm requires some special features if the animal is previously known. Still, the animal images are also manually cropped to select only the whole animal shape before feeding to the conventional machine learning algorithm.

Deep learning approaches have been shown to supersede traditional machine learning algorithms in vision tasks. Researchers who used deep learning approaches used their datasets with one or a few animal species (Big animals like Lion,

elephants, etc.), and others used relatively small datasets (a few thousand images) only. In contrast, according to the literature we found, others relied on feature extraction descriptors to classify animals. On the other hand, attention mechanisms on deep learning approaches have not been examined by researchers for the task of animal species and detection as techniques based on attention mechanisms emphasize crucial information and disregard inconsistency. Thus this led to the proposal of Animal Species Detection and Classification Framework based on modified multi-scale Attention Mechanism and Feature Pyramid Network using Animal 80 dataset and Animal Wildlife animal as shown in Figure 5.

Materials and methods

Proposed architecture

To observe and identify different kinds of animal species, we utilized a two-stage network illustrated in Figure 1. Following the work of [50]· we used the backbone network topology of ResNet50. We applied a modified multiscale attention mechanism to the backbone network to focus on the features maps for generating a future pyramid. The modified multiscale attention model is processed in five scales designated $\{c_1, c_2, c_3, c_4, c_5\}$ before the feature pyramid is gained from the feature maps to produce high-level semanticized maps at any scale. The main aim is to select the fine details of feature maps depending on contextual features from top-down levels and pick contextual factors from top-down depending on tiny low-level data. We used a collection of pyramidal features to obtain a feature pyramid with robust semantics on all scales from the individual phases of the multiscale processed ResNet-50 features. The pyramid path utilizes lower standards to recognize tiny animal species, and the vertical link contributes to strengthening connections between low and high-level elements. In ResNet50, c_1 is the function output of the conv1x layer, c_2 is the res2bx layer's output of the feature, c_3 is the Res3ax layer's output, c_4 is res4b22x layer's result, and C_5 is the res5cx layer feature output. Input to the RPN back layer by layer is all outputs of multiscale attention models. In particular, c_{i-1}'s block, $p_{i-1,}$ developed with location data processed by P_{i-1}, is used for training the RPN. The $\{p_2, p_3, p_4, p_5\}$ characteristics maps produced by a pyramid network have high-resolution geographical information, and low-resolution semanticized information. The last function map is mixed with the preceding function maps, then sent via RPN, and the detection output is finally achieved. We employed the DenseNet model [48] in the classification phase.

DenseNet is more efficient but decreases the number of parameters when merging feature mappings from the varying levels, increasing variations in future layers' input. We also used focal Loss [1] to keep track of our network to improve precision.

Attention mechanism

Following the implementation of multiscale attention [50] depicted in Figure 2, to enhance detection accuracy in complex real-world scenarios and enhance the range to locate the hard-negative samples, we modified the multiscale attention for animal species detection and Classification. Dot-product multiplications, weight gains, and SoftMax are achieved through the process, resulting in the animal species characteristics when summed up, thereby enhancing the detection accuracy. Given the same multiple function map by feature pyramid *Q, K, V,* the dot product is carried out between Q^T *and* K to get their similitude measures as follows;

$$f\left(Q^T,\ K_j\right),\ j = 1,\ 2,\ \ldots,\ m \tag{1}$$

Where $f(.)$ represents the dot-product operation, j represents the j^{-th} feature map. We achieved the normalization s_j by;

$$s_j = Softmax\left(\frac{f\left(Q^T,\ K\right)}{\sum_{j=1}^{m} e^{f\left(Q,\ K_j\right)}}\right),\ j = 1,\ 2,\ \ldots.,\ m \tag{2}$$

After attention processing, the final characteristic map is fused into each layer, thus yielding the modified multiscale Attention defined as;

$$Attention(Q,\ K,\ V) = Softmax\left(\frac{f\left(Q^T,\ K\right)}{\sum_{j=1}^{m} e^{f\left(Q,\ K_j\right)}}\right)V \tag{3}$$

Finally, the multiscale attention vector is obtained by adding all the weights thus;

$$\sum_{i=1}^{m} s_i V_i,\ \ (m = 4 = \ number\ of\ added\ weights) \tag{4}$$

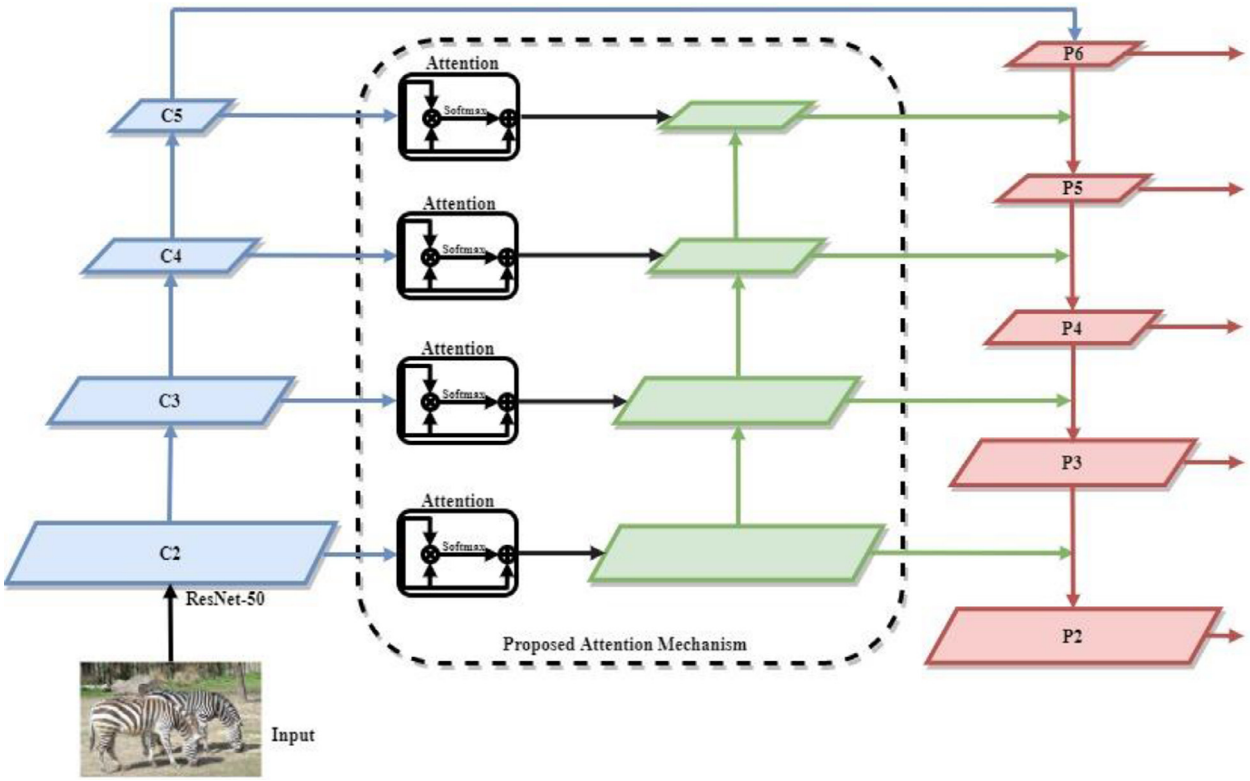


Figure 2. Modified Multi-scale Attention Mechanism

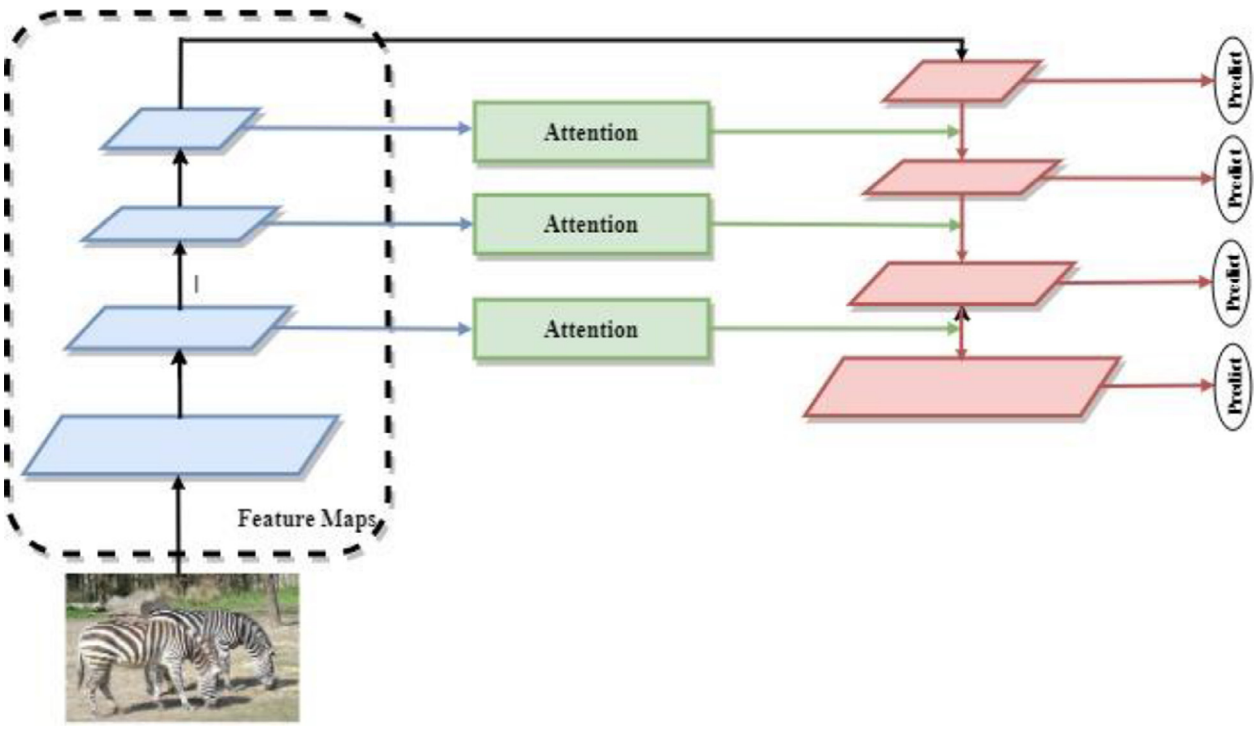


Figure 3. Illustration of the Feature Pyramid Network

Pyramidal feature network

We built the Pyramidal Feature Networks (PFN) in the region proposal stage. The Feature Map (FM) size will differ after going through different convolutional layers in the backbone network ResNet-50 [42] and passing through the multiscale attention mechanism, as seen in Figure 3. This FM is extracted in varying sizes, such that it can make a characteristic pyramid and is referenced as $\{c_2, c_3, c_4, c_5\}$. The products of attention models are all entered layer by layer into the Feature pyramid. We also see that most tiny animal species are missing when only $\{c_3, c_4, c_5\}$ is used. If just $\{c_2, c_3, c_4\}$ are used, it cuts most large animal species. Just as we know, the convolutional layers merge the characteristics and expand the receptive field to use low-level detection features.

Adding to the attention model, FPN increases the low resolution, unanticipatedly strong characteristics from the top with the high resolution, unanticipatedly weak features from the bottom. Therefore, creating a multi-scale feature pyramid with rich semantic representation at all levels is straightforward and practical. We do this till the lowest feature level fuses with the most significant level. Creating the network pyramid architecture is clear, leading to high-level semantic characteristic maps at every scale, although with minor extra costs.

The technique used to integrate all levels is to upgrade the more abstract and semantically strengthened higher-level feature mappings and then laterally link the elements to the last layer, increasing the higher-level characteristics. It should be noticed that the two layers of laterally related aspects must be the same in space dimensions as they must be concatenated.

Architecture details

We used 2 as it is the closet neighbor upsampling method to enhance the high-level feature and then concatenate with the respective preceding layer feature (the previous layer would be put through to a $1 * 1$ convolution kernel, and the aim is to change the channels that must be the same as the channels in the next layer). This approach is iteratively carried out until the most refined feature map is created. A $1 * 1$ convolutional kernel is added when the iteration initializes to make the roughest feature map behind the c_5. Finally, we employ a $3 * 3$ convolution kernel to create the last needed feature map, so the aliasing effect of a sample up may be eliminated. The respective levels of features $\{c_2, c_3, c_4, c_5\}$ are $\{p_2, p_3, p_4, p_5\}$, and the space layer remains the same. From Fig. 3, we find out that the lowest characteristic map called c_5 is utilized twice to refine the semantic image of the small item. Therefore, eventually, several character maps may be obtained as $\{p_2, p_3, p_4, p_5, p_6\}$, equivalent to $\{c_2, c_3, c_4, c_5\}$.

ROIs are essential for the entire procedure; therefore, we need the anchors for tiny object identification to be appropriately designed. The RPN anchors [5] are boxes having pre-defined proportions of sizes and aspects. We initiate anchors with $\{8^2, 16^2, 32^2, 64^2, 128^2\}$ pixels to recognize tiny animal species. All the markings do not face the camera directly; therefore, we set up many aspect ratios on every level using {1:2; 1:1; 2:1} and dimensional anchors $\{2^0, 2^{1/3}, 2^{2/3}\}$. Thus, the pyramidal feature hierarchy has $3 * 3 * 5 = 45$ anchors. The anchor steps are altered to {4, 8, 16, 32, 64} to adjust to reduced dimensional characteristics. For detection, these four outputs for each anchor per space forecast the relative compensation between the anchor and the ground-truth box at each level of the pyramid feature at several scales.

Loss function

We have used the Focal Loss [1] to resolve the minimal class imbalance challenge during training with remarkable success. Focal Loss is useful where the call imbalance is minimal, as in our Animal Wildlife dataset. When applied with data augmentation, the class imbalance of training is always corrected. In this paper, we only used the Focal Loss as a new loss function to make the tough cases contribute more to the Loss and help the network learn from these tricky examples. We define *FL* as mathematically;

$$FL(p_t) = -\alpha_t (1 - p_t)^y \log(p_t) \tag{5}$$

Where p_t is defined as;

$$p_t = \begin{cases} p & if\ y = 1 \\ 1 - p & otherwise \end{cases} \tag{6}$$

The focus loss concentrates on a few challenging instances and successfully reduces the effects of simple examples.

Dataset and data preprocessing

This paper experimented on two animal datasets: the African Wildlife dataset and the Animal-80 dataset. Both datasets are a single class in a scene.

Animal-80 dataset: This consists of 80 classes of animal species, with 45,132 for the training set and 13,010 for the test set. Considering that the employed dataset is not robust enough, some categories such as the Bull, canary, seahorse, shrimp, squid, and turtle classes have few instances with 94, 42, 14, 60, 30, and 48 instances, respectively. The repository for this dataset can be found in https://www.kaggle.com/antoreepjana/animals-detection-images-dataset.

Animal Wildlife: This dataset consists of four categories of animal species, namely Buffalo, Elephant, Rhino, and zebra, with each class having 752 instances with a total number of 3,008 images. Because of the variance in the size of the employed dataset, we resized it to 356*356 and fed it into our model. We only recomputed the annotation after resizing since the Animal-80 dataset was in the pascalVoc annotation file as a data preprocessing step already. The African

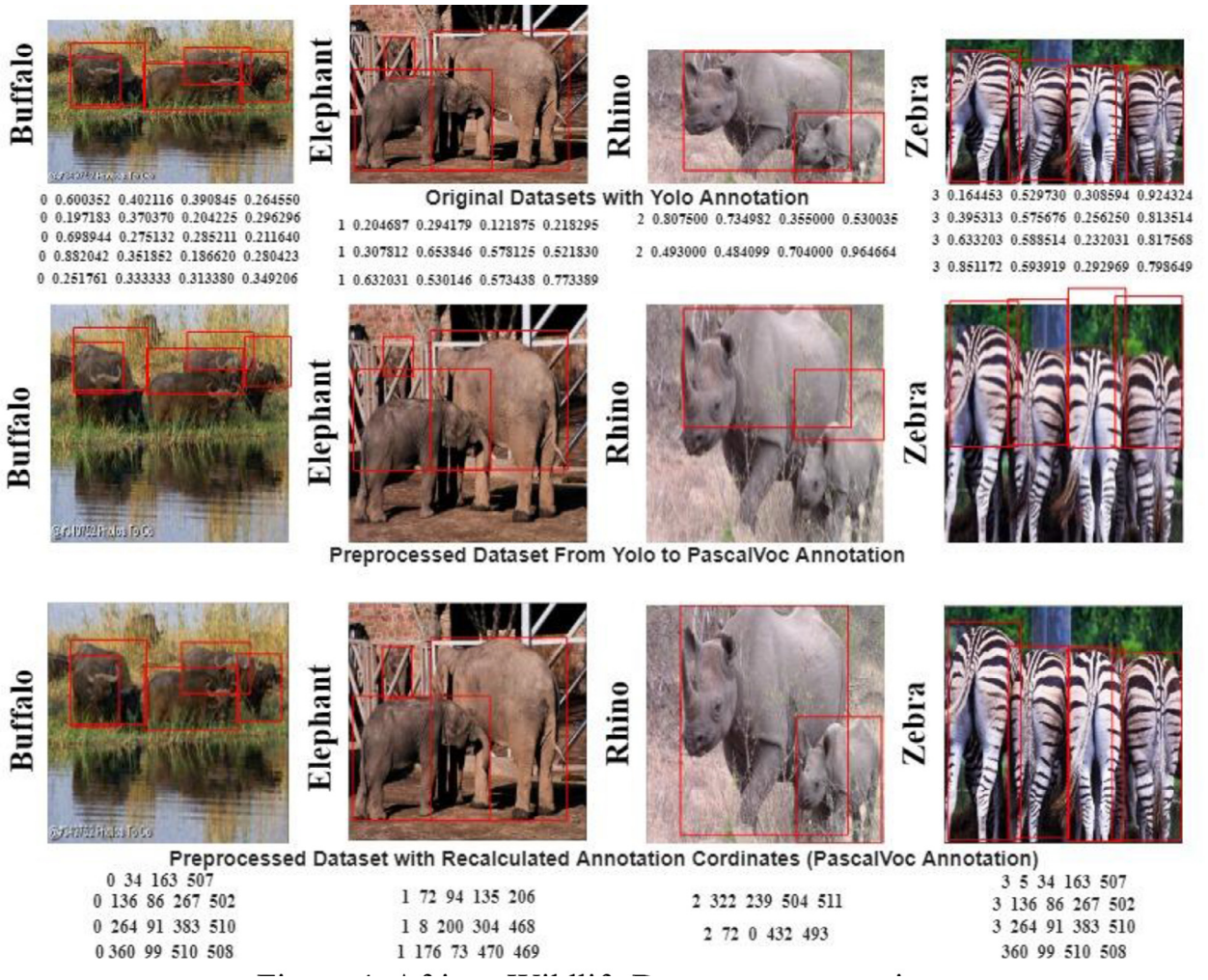


Figure 4. African Wildlife Dataset preprocessing

wildlife dataset is initially in a Yolo annotation format, as depicted in Figure 4, which forced us to develop a simple conversion python code to do both resizing, reannotation, and conversion from Yolo to PascalVoc annotation file format. https://www.kaggle.com/biancaferreira/african-wildlife gives access to the dataset.

Evaluation metrics

We used the Average Precision (AP) and Mean Average Precision (mAP) metrics to evaluate our proposed method, which is mathematically defined as;

$$AP = \int_0^1 P(r)dr \tag{7}$$

In Eq. 7, P denotes precision calculated mathematically as; $P = \frac{st_p}{st_p + sf_p}$, where st_p denotes true positive (TP), indicating that the detections of animal species are correct while sf_p Denotes false positive (FP), indicating the false detection of the animal species. r represents recall and is calculated thus; $P = \frac{st_p}{st_p + sf_n}$, where sf_n Represents false negative, which means unidentified animal species.

$$mAP = \frac{1}{s}\int_0^1 P(r)dr \tag{8}$$

Where s represents the total number of animal species

Experiment

The experimental evaluation of our proposed algorithm was carried out and identified the effect of attention mechanism in our techniques in this section. We used a python deep learning open source framework deployed in the Keras-TensorFlow code to implement our Animal species detection and classification algorithm. We evaluated our techniques using the African Wildlife animal dataset and Animal-80 dataset.

Experimental setup

We experimented on a python environment with Intel® Core™ i7-7700k CPU @ 4.20GHz CPU of 32.0 GB, 64-bit Operating System, x64-based processor, and NVIDIA GeForce GTX 1080TI GPU (12GB Memory). The Resnet-50 architecture, which is pretrained on the ImageNet1k dataset, is applied as our backbone, as depicted in Fig.2. Changing high-resolution images to low-resolution images, which suit the anchor sizes, improves detecting large Animal species. We set $\gamma = 2$ and $\alpha_t = 0.25$ in Eq. (5). The parameters of our architecture: we trained with stochastic gradient descent (SGD) and set weight decay to 0.0001, momentum to 0.9, learning rate =1e-4, epoch = 100, batch size =8, reduced learning rate factor=0.1 we used a single GPU while training.

Results and discussions

This section reports the experimental results. We highlight the results of our proposed model on the evaluation datasets separately for fair judgment.

African wildlife dataset

Figure 6. shows the average precision of each animal class: Rhino, Zebra, Buffalo, and Elephant. We display the AP performance on both our attention and without attention mechanism.

A. Attention B. Without Attention

When our attention module is embedded in our network depicted in Figure 6(A.), the model outperforms the instance without attention described in Figure 6(B.) in terms of average precision. Additionally, we could see that the with-attention network yields a smoother learning curve than the without-attention network, which we can see from the *elephant* animal class. Similar cases are observed in the other classes.

A. Attention B. Without Attention

The losses (Figure 7) also show that the with-attention network exhibits a smoother learning process than the without-attention network. Diving into the detection outcomes, from Figure 8, we could see that the without attention network identified the *human* in the image as an *elephant.* In contrast, the with-attention network ignored the *human* because it is not part of targeted detection objects. Generally, Figure 9(A.) indicates that the with-attention network significantly outperforms the without-attention network in terms of maP with a margin of about 0.02%. This proves the relevant role attention mechanism can play to increase deep learning network performance.

Animal-80 dataset

This section talks about the results obtained from the evaluation of the African dataset. Because of the large samples of this dataset, results demonstrate how promising the with-attention is. From Figure 10, the with-attention network outperforms the without-attention network by obtaining a minimal loss in regression, Classification, and mean squared error loss.

A. African Wildlife Dataset B. Animal-80 Dataset

Moreover, compared to maP, the with-attention network significantly surpasses the without-attention network with a margin of about 0.1%, as depicted in Figure 9B. This, without doubt, validates the effectiveness of our proposed network with attention Figures 12. and 13 show the best maP achieved for each animal class at a particular epoch.

A. With Attention B. Without Attention

From Figure 11. We can identify the penalty of dataset imbalance on the detection performance for the Animal-80 dataset. While training supported by the AP results, we noticed that classes with larger images performed better than those with few images since we didn't carry out data augmentation during our data preprocessing.

Discussion

In this paper, we presented a modified multi-scale attention and feature pyramid-based deep learning framework for Animal species detection and Classification, a significant challenge in the research community due to the associated animal features such as size, colour, shape, etc. We used the dataset as downloaded without changing the training and testing set partitions. We only resized the datasets, and variation in size was seen in various classes of each dataset to 356 *356. Due to the resizing, there was a need for annotation recomputation for the Animal wildlife dataset and conversion and recomputation of the Animal-80 dataset annotation from YOLO format to Pascal VOC as illustrated in the figure. 4. The

Table 1
Summary of African Wildlife Dataset of the AP @ Epoch 50, Epoch 100 and the maP at the Corresponding Epochs.

African Wildlife Dataset	$AP_{Attention}$ @ Epoch50	$AP_{No\text{-}Attention}$ @ Epoch50	$maP_{Attention}$ @ Epoch50	$maP_{No\text{-}Attention}$ @ Epoch50	$AP_{Attention}$ @ Epoch100	$AP_{No\text{-}Attention}$ @ Epoch100	$maP_{Attention}$ @ Epoch100	$maP_{No\text{-}Attention}$ @Epoch100
Rhino	0.92	0.90	0.87	0.85	0.93	0.89	0.92	0.90
Zebra	0.88	0.86			0.92	0.87		
Buffalo	0.86	0.84			0.89	0.85		
Elephant	0.82	0.80			0.88	0.82		

datasets we used have only one animal class per scene. We implemented two different types of methods: the modified multiscale attention mechanism and the conventional proposed architecture. The animal wildlife classes are almost equal; thus no need for data argumentation. The best AP values are seen between epoch 25 and 40, where the individual AP's reaches the maximum and diminish as the epoch goes to 100. This is a clear indication of Training parameters not duly set, which we will look into as future work. As seen in Table 4 during the training, the classes with very few instances had low AP. Compared with the conventional model, it is noted that when our modified multiscale attention is trained with a balanced dataset, it shows a high performance than when trained with an imbalance set. However, it increases the AP with exceptional values, as seen in table 4 Tables 1. and 2 summarize the AP and maP of the implemented algorithms at epoch 50 and 100, respectively.

Comparison with other algorithms using the african wildlife dataset

Since the African Wildlife Dataset contains only four classes without the issue of imbalance in the dataset, we used the source codes provided by [38]. We evaluated the African Wildlife Dataset on them for a fair comparison as shown in Table 5. They implemented several models such as YOLOv3 [59], which is a one stage object detection network with a high-speed image processor and slightly lower quality, RetinaNet R-50-FPN[1], which is also a one-stage object detection architecture but designed mainly to test the Focal Loss function that was introduced for training accuracy improvement, Faster R-CNN R-50-FPN [5] which is a two-stage architecture that uses the ResNet-50 architecture with FPN for feature extraction and lastly, Cascaded R-CNN R-50-FPN [46] which is a multi-stage object detection algorithm. We also used the Dataset (10 Classes) recorded in their paper to validate our model's superiority further.

Limitations and future works

From table 3, for the Animal-80 Dataset, we saw the low AP of various animal species with few instances which shows that we did not consider data augmentation during our experiments. The modified attention mechanism-sensitive parameters were not considered. We did not view images of different classes in a scene in our experiment due to the experimented dataset. For future work, it is necessary to address the below-listed limitations.

- Experimenting with a dataset with multiple classes in a scene.
- Analysis of the sensitive parameters of the modified multiscale attention mechanism for the best and worst-case scenarios.
- We will consider Data augmentation for the Animal-80 dataset, compare the modified attention mechanism, and use other networks.

Conclusion

This paper proposed a novel algorithm for animal species detection and Classification. In the first detection step of our algorithm, aiming at solving the problem of detecting animal species caused by the size, shape, colour as well as the changeable natural environment and behavior of an animal, we use the ResNet-50 as the backbone to build a multiscale attention mechanism and feature pyramid network which can adequately enhance the detection of small animal species. We extracted the feature maps using the network backbone enhanced by the modified multiscale attention before passing it into the future pyramid designated $\{c_1, c_2, c_3, c_4, c_5\}$ in five scales to produce high-level semanticized maps at any scale. The main aim is to select the fine details of feature maps depending on contextual features from top-down levels and pick contextual factors from top-down depending on tiny low-level data. We achieved a similarity measurement by a dot-product multiplication and softmax with the input of the multiscale attention features; therefore, we weighed and summed the features to enhance the detection accuracy. To further improve the performance of our network, this paper employs the use of the Focal Loss to supervise the region proposal network. In the second step (recognition), we used the dense blocks to enhance the accuracy of the classification network. The proposed model was evaluated using AP and Map evaluation metrics on the African wildlife dataset and the challenging Animal-80 dataset. The result demonstrates that the proposed method performs better with the attention mechanism's support. In the future, our primary focus will be on addressing the noted limitations

Table 2
Animal-80 Dataset Results Summary of the AP @ Epoch 50, Epoch 100 and the best AP the Corresponding Epochs.

Animal-80 Dataset	$AP_{Attention}$ @ Epoch50	$AP_{No\text{-}Attention}$ @ Epoch50	$AP_{Attention}$ @ Epoch100	$AP_{No\text{-}Attention}$ @ Epoch100	Best $AP_{Attention}$	@Epoch	Best $AP_{No\text{-}Attention}$	@Epoch
Bear	0.12	0.16	0.11	0.16	0.19	37	0.18	85
	0.24	0.07	0.17	0.11	0.34	25	0.17	15
Brown_Bear								
Bull	0.10	0.10	0.07	0.08	0.17	16	0.13	27
Butterfly	0.84	0.81	0.83	0.82	0.89	26	0.85	13
Camel	0.21	0.09	0.14	0.10	0.29	25	0.15	58
Canary	0.54	0.30	0.56	0.32	0.63	57	0.45	23
Caterpiller	0.62	0.45	0.60	0.48	0.65	51	0.53	22
Cattle	0.26	0.12	0.21	0.10	0.3	23	0.21	17
Centipede	0.67	0.52	0.70	0.50	0.72	31	0.55	21
Cheetah	0.39	0.30	0.38	0.30	0.43	40	0.33	64
Chicken	0.83	0.67	0.81	0.68	0.81	24	0.70	59
Crab	0.51	0.58	0.51	0.55	0.53	53	0.61	34
Crocodile	0.66	0.60	0.64	0.60	0.72	37	0.64	23
Deer	0.68	0.56	0.67	0.48	0.71	19	0.60	23
Duck	0.36	0.37	0.33	0.38	0.40	55	0.40	86
Eagle	0.88	0.83	0.88	0.86	0.90	25	0.87	85
Elephant	0.32	0.267	0.29	0.32	0.41	23	0.41	23
Fish	0.42	0.31	0.38	0.29	0.46	18	0.36	13
Fox	0.78	0.51	0.77	0.50	0.83	21	0.57	23
Frog	0.63	0.63	0.61	0.61	0.67	40	0.66	37
Giraffe	0.76	0,71	0.78	0.70	0.79	75	0.71	48
Goat	0.25	0.25	0.25	0.24	0.30	34	0.25	66
Goldfish	0.10	0.10	0.10	0.10	0.13	40	0.11	73
Goose	0.26	0.18	0.22	0.20	0.32	28	0.23	61
Hamster	0.22	o.18	0.15	0/10	0.30	18	0.21	28
	0.30	0.26	0.28	0.27	0.36	20	0.30	74
Harbor_Seal								
Hedgehog	0.84	0.77	0.81	0.75	0.87	28	0.30	64
	0.31	0.22	0.35	0.21	0.46	22	0.80	23
Hippopotamus								
Horse	0.40	0.32	0.36	0.31	0.41	26	0.30	51
Jaguar	0.34	0.28	0.34	0.30	0.39	40	0.35	55
Jellyfish	0.68	0.47	0.67	0.43	0.70	25	0.35	20
Kangaroo	0.35	0.16	0.28	0.17	0.38	48	0.56	72
Koala	0.30	0.10	0.25	0.12	0.34	59	0.31	77
Ladybug	0.90	0.77	0.90	0.77	0.91	65	0.80	44
Leopard	0.46	0.42	0.45	0.40	0.48	63	0.45	64
Lion	0.80	0.77	0.78	0.75	0.84	18	0.82	22
Lizard	0.65	0.88	0.63	0.87	0.72	28	0.90	31
Lynx	0.53	0.49	0.57	0.44	0.62	22	0.55	50
Magpie	0.85	0.54	0.84	0.49	0.88	21	0.56	68
Monkey	0.39	0.81	0.36	0.80	0.42	14	0.82	66
Moths and Butterflies	0.31	0.37	0.28	0.37	0.35	37	0.44	13
Mouse	0.38	0.27	0.37	0.26	0.42	33	0.32	54
Mule	0.64	0.22	0.62	0.22	0.71	18	0.27	38
Ostrich	0.39	0.36	0.35	0.48	0.44	21	0.54	58
Otter	0.83	0.27	0.82	0.31	0.87	23	0.40	35
Owl	0.54	0.23	0.52	0.85	0.61	24	0.86	81
Panda	0.71	0.49	0.70	0.42	0.76	18	0.48	74
Parrot	0.48	0.33	0.46	0.64	0.48	61	0.68	55
Penguin	0.36	0.84	0.34	0.42	0.40	16	0.50	45
Pig	0.82	0.40	0.82	0.37	0.85	34	0.44	74
Polar_Bear	0.70	0.67	0.65	0.65	0.72	16	0.75	38
Rabbit	0.49	0.43	0.44	0.48	0.53	35	0.54	36
Raccoon	0.49	0.44	0.44	0.45	0.56	33	0.54	37
Raven	0.48	0.43	0.52	0.38	0.57	64	0.48	59
Red Panda	0.65	0.46	0.61	0.49	0.65	49	0.51	89
Rhinoceros	0.58	0.60	0.53	0.59	0.60	46	0.63	68
Scorpion	0.80	0.77	0.70	0.64	0.80	18	0.77	49
Sea_Lion	0.24	0.21	0.20	0.19	0.35	17	0.24	27
Sea_Turtle	0.30	0.23	0.24	0.14	0.37	10	0.28	11
Seahorse	0.08	0.08	0.08	0.08	0.13	25	0.12	77
Shark	0.52	0.405	0.48	0.37	0.54	33	0.45	37

(*continued on next page*)

Table 2 (*continued*)

Animal-80 Dataset	$AP_{Attention}$ @ Epoch50	$AP_{No-Attention}$ @ Epoch50	$AP_{Attention}$ @ Epoch100	$AP_{No-Attention}$ @ Epoch100	Best $AP_{Attention}$	@Epoch	Best $AP_{No-Attention}$	@Epoch
Sheep	0.08	0.02	0.06	0.03	0.13	25	0.05	15
Shrimp	0.11	0.08	0.08	0.05	0.28	4	0.18	8
Snail	0.78	0.70	0.77	0.70	0.83	23	0.72	33
Snake	0.90	0.90	0.90	0.88	0.91	40	0.90	48
Sparrow	0.83	0.87	0.81	0.86	0.88	21	0.90	82
Spider	0.92	0.91	0.91	0.90	0.94	17	0.94	18
Squid	0.003	0.003	0.01	0.01	0.03	8	0.09	87
Squirrel	0.50	0.40	0.50	0.41	0.56	22	0.45	66
Starfish	0.80	0.78	0.80	0.80	0.86	16	0.82	87
Swan	0.47	0.37	0.43	0.36	0.50	26	0.43	24
Tick	0.33	0	0.33	0.02	0.50	15	1	31
Tiger	0.74	0.62	0.75	0.60	0.80	52	0.67	60
Tortoise	0.50	0.47	0.50	0.46	0.51	58	0.50	47
Turkey	0.50	0.22	0.50	0.28	0.52	59	0.33	57
Turtle	0	0	0	0	0.01	15	0.003	7
Whale	0.62	0.52	0.62	0.51	0.63	28	0.56	77
Woodpecker	0.62	0.48	0.62	0.60	0.64	48	0.60	94
Worm	0.20	0.18	0.21	0.17	0.26	46	0.25	82
Zebra	0.67	0.63	0.64	0.60	0.72	21	0.66	6

The summary of the validation regression loss, classification loss, and Loss is presented in Table 3. We noted that the best Loss is at a different epoch which could be that the learning rate is high and when it's reaching the local minimum, it exceeds it, or at some point, the invalid probabilities are being passed to the loss function.

Table 3
Validation Losses Summary

African Wildlife Dataset	Attention			No_Attention		
	Regression Loss	**Classification Loss**	**Loss**	**Regression Loss**	**Classification Loss**	**Loss**
Epoch @25	0.69	0.20	0.89	0.72	0.21	0.92
Epoch @50	0.62	0.32	0.94	0.65	0.34	1.0
Epoch @75	0.62	0.44	1.16	0.65	0.52	1.18
Epoch @100	0.62	0.52	1.14	0.65	0.52	1.18
Best	0.62@17	0.19@24	0.89 @24	0.64 @43	0.18 @26	0.92 @25
Animal-80 Dataset	**Attention** **Regression Loss**	**Classification Loss**	**Loss**	**No_Attention** **Regression Loss**	**Classification Loss**	**Loss**
Epoch @25	0.54	0.57	1.11	0.62	0.60	1.22
Epoch @50	0.58	0.95	1.53	0.66	1.07	1.73
Epoch @75	0.58	1.19	1.77	0.65	1.50	2.15
Epoch @100	0.59	1.54	2.13	0.66	1.58	2.24
Best Epoch@7	0.52 **@15**	0.43 **@7**	0.90 **@15**	0.57 **@7**	0.39 **@15**	1.01 **@7**

Table 4
AP Comparison Result on the 10 Classes Dataset used in [38] @ Epoch 50

	Cascaded R-CNNR-50-FPN	Faster R-CNNR-50-FPN	RetinaNetR-50-FPN	YOLOv3	OursNo Attention	OursWith Attention
Dog	0.81	0.81	0.83	0.92	0.85	0.90
Horse	0.75	0.76	0.77	0.88	0.80	0.89
Sheep	0.68	0.67	0.65	0.75	0.70	0.76
Cow	0.65	0.66	0.60	0.80	0.72	0.80
Elephant	0.82	0.83	0.84	0.88	0.88	0.90
Bear	0.81	0.87	0.89	0.95	0.90	0.94
Zebra	0.84	0.88	0.88	0.91	0.90	0.92
Giraffe	0.87	0.86	0.87	0.91	0.89	0.93
Fox	0.21	0.18	0.19	0.18	0.20	0.23
Goat	0.39	0.44	0.41	0.58	0.47	0.60
maP	**0.68**	**0.70**	**0.69**	**0.78**	**0.73**	**0.79**

From Table 4, the YOLOv3 outperformed our proposed model with a slight difference in some of the classes when implemented on the 10 classes dataset recorded in [38]. In comparison, our algorithm outperformed all other compared architectures on the African Wildlife Dataset with 4 classes. During training, we observed that tunning our model parameters such as Learning Rate, Batch Size, Epoch, Weight Initialization, Activation Functions, etc., increases detection accuracy.

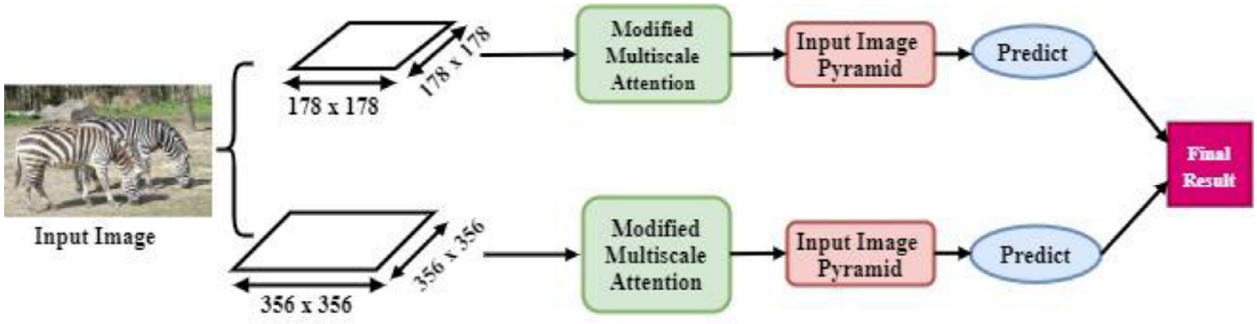
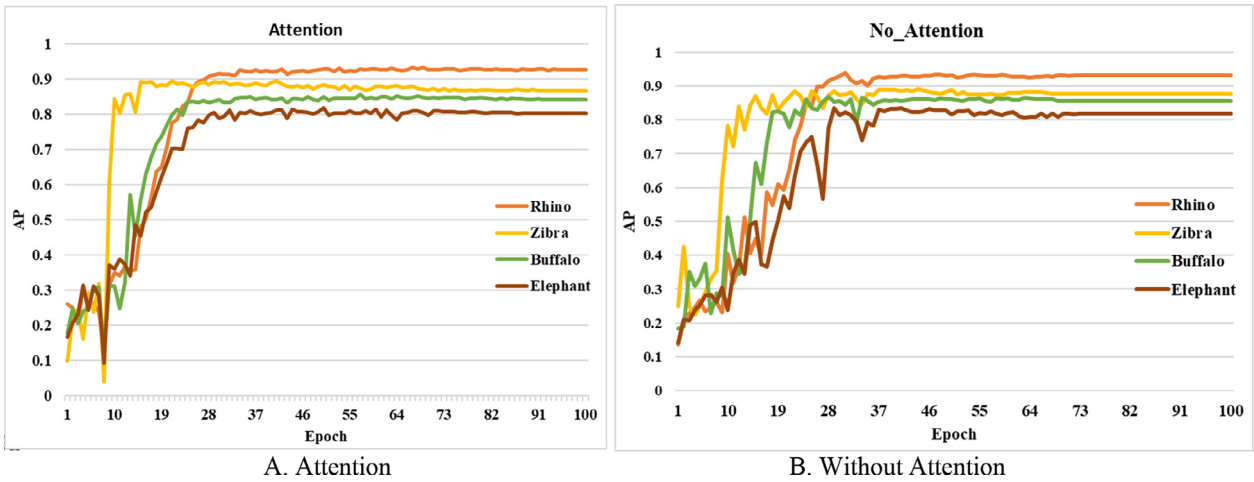
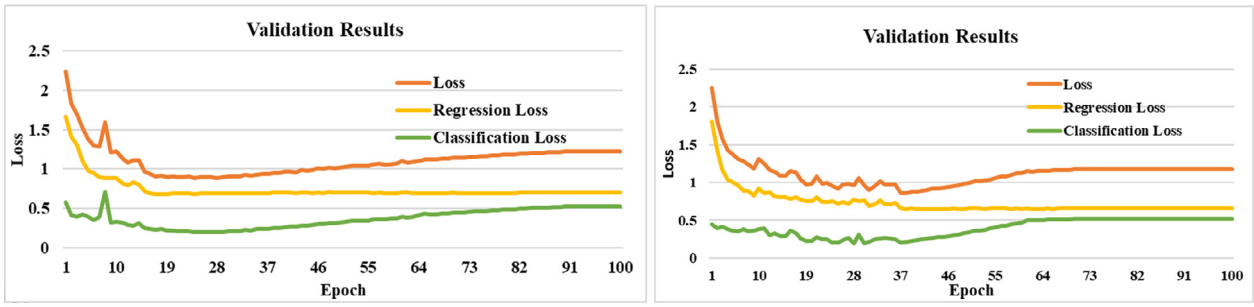


Figure 5. Illustration of our implementation details. The original input image is resized to its half before feeding both into our model to get the final prediction.



A. Attention B. Without Attention

Figure 6. Average Precision on African Wildlife Dataset



A. Attention B. Without Attention

Figure 7. Validation Result on African Wildlife Dataset

to make the model more robust and vast. Carrying out a sensitivity parameter setting test on the proposed model with a dataset of multiple classes in a scene while applying data augmentation will make the proposed model more suitable for animal detection and classification and computer vision problems in general.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

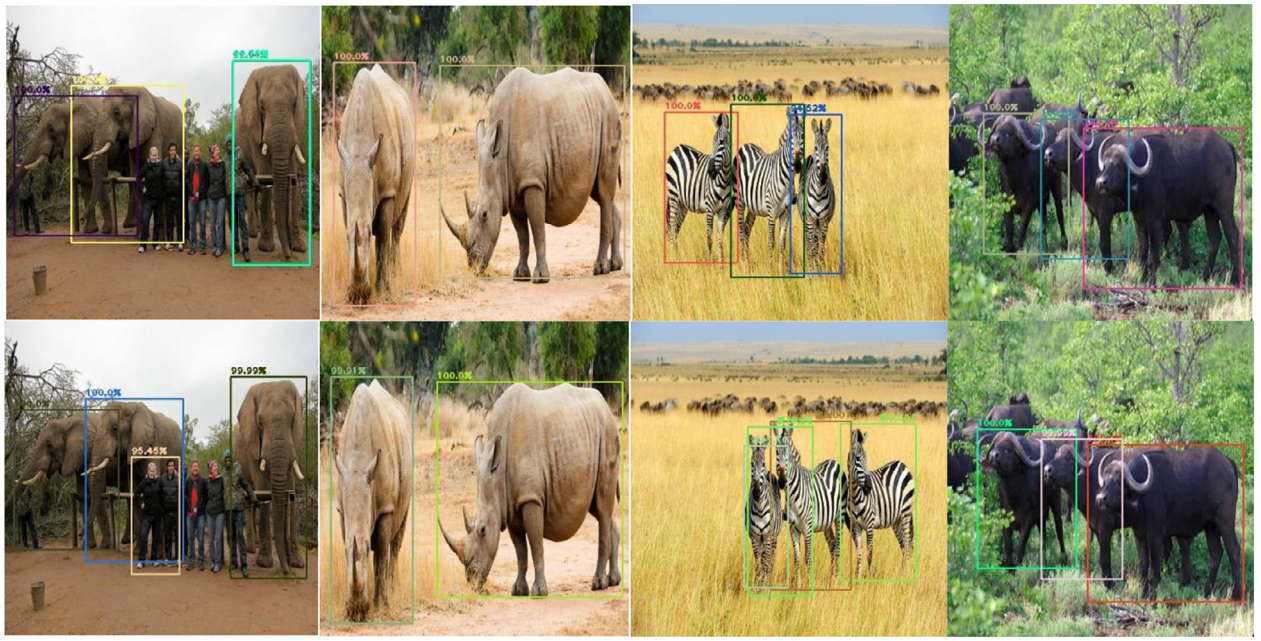


Figure 8. Recognition results. The upper row represents the attention mechanism, while the lower row illustrates the without attention mechanism.

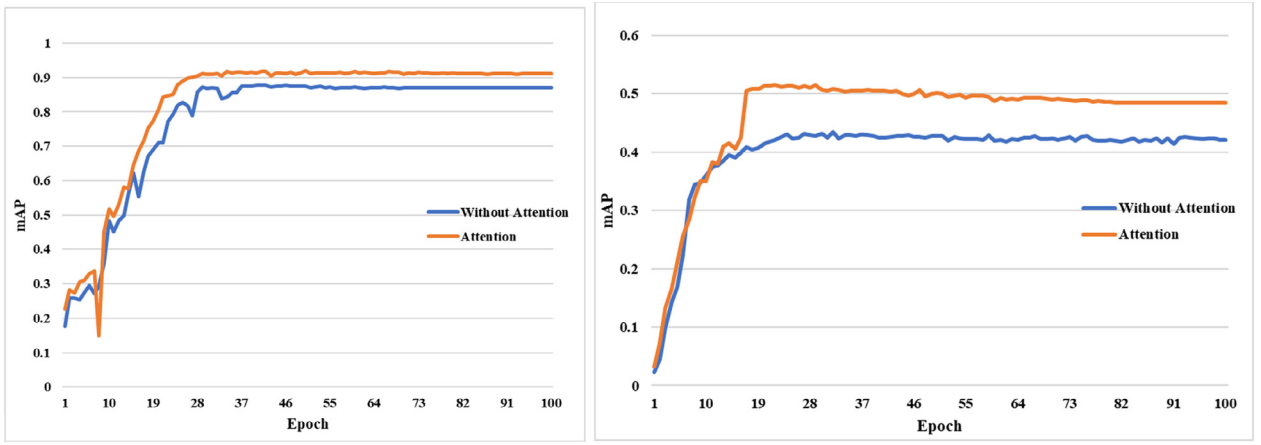


Figure 9. maP Result

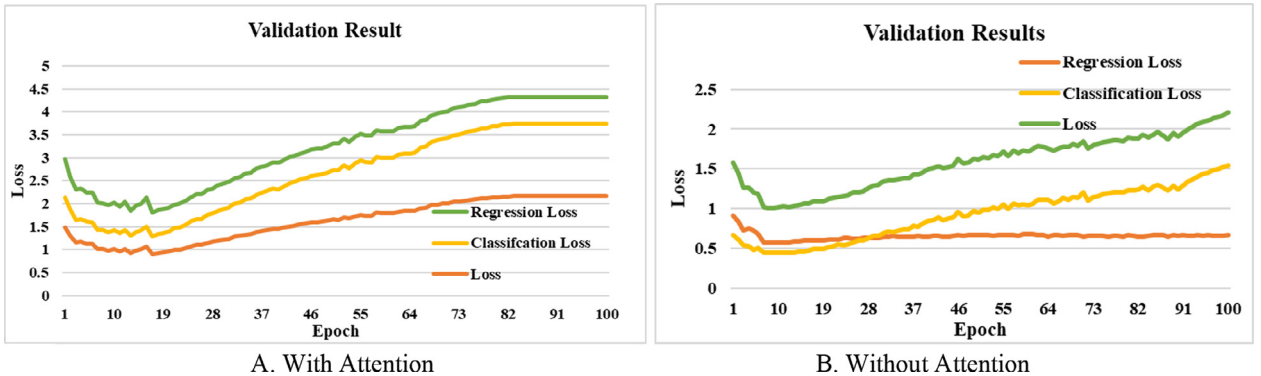


Figure 10. Validation losses on Animal 80 Dataset

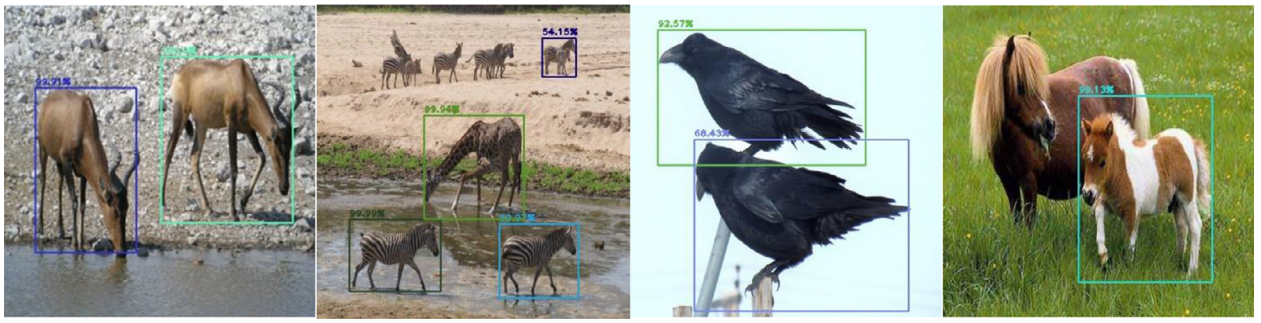


Figure 11. Detection Performance on the Animal-80 Dataset

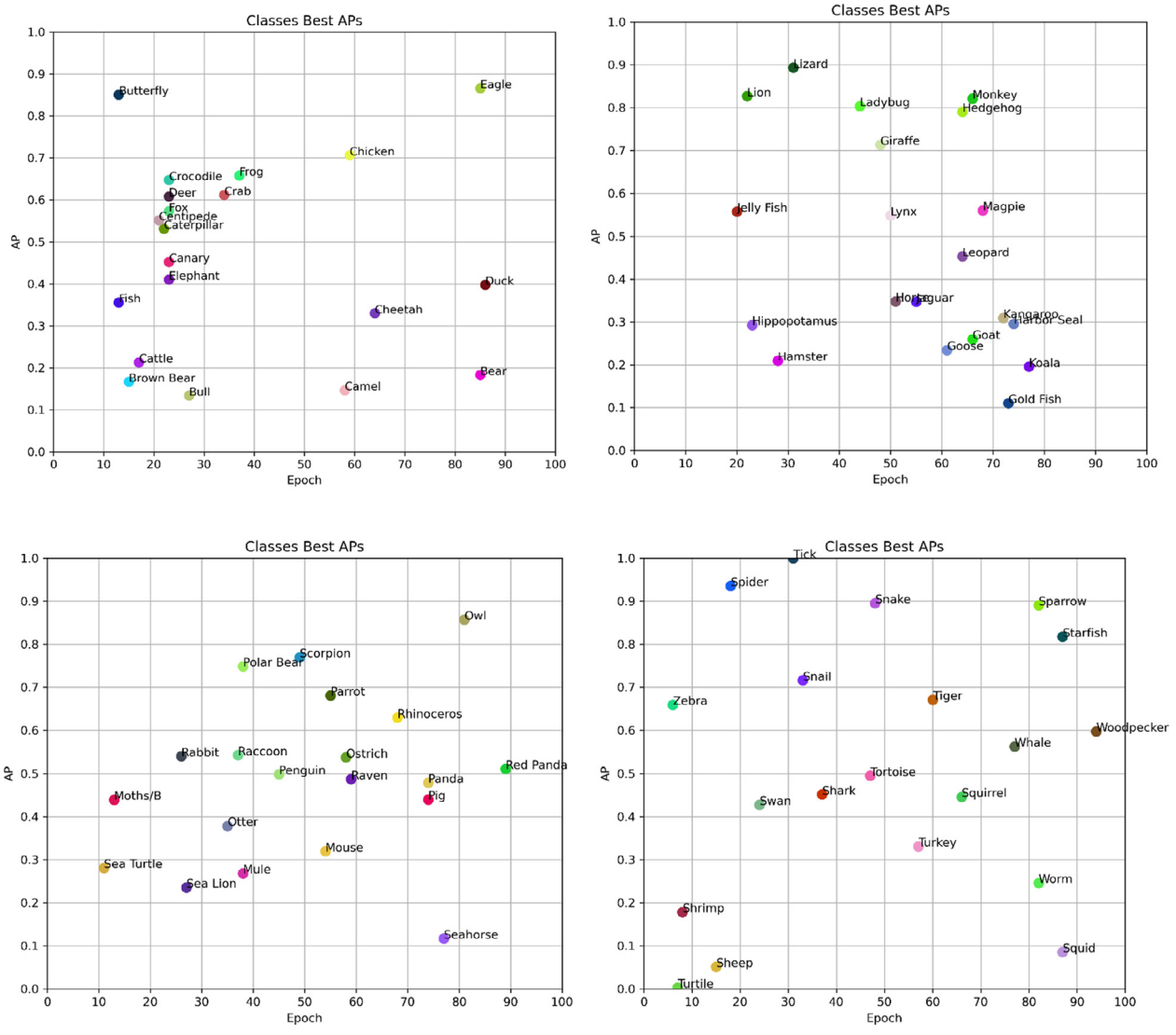


Figure 12. Average Precision Attention

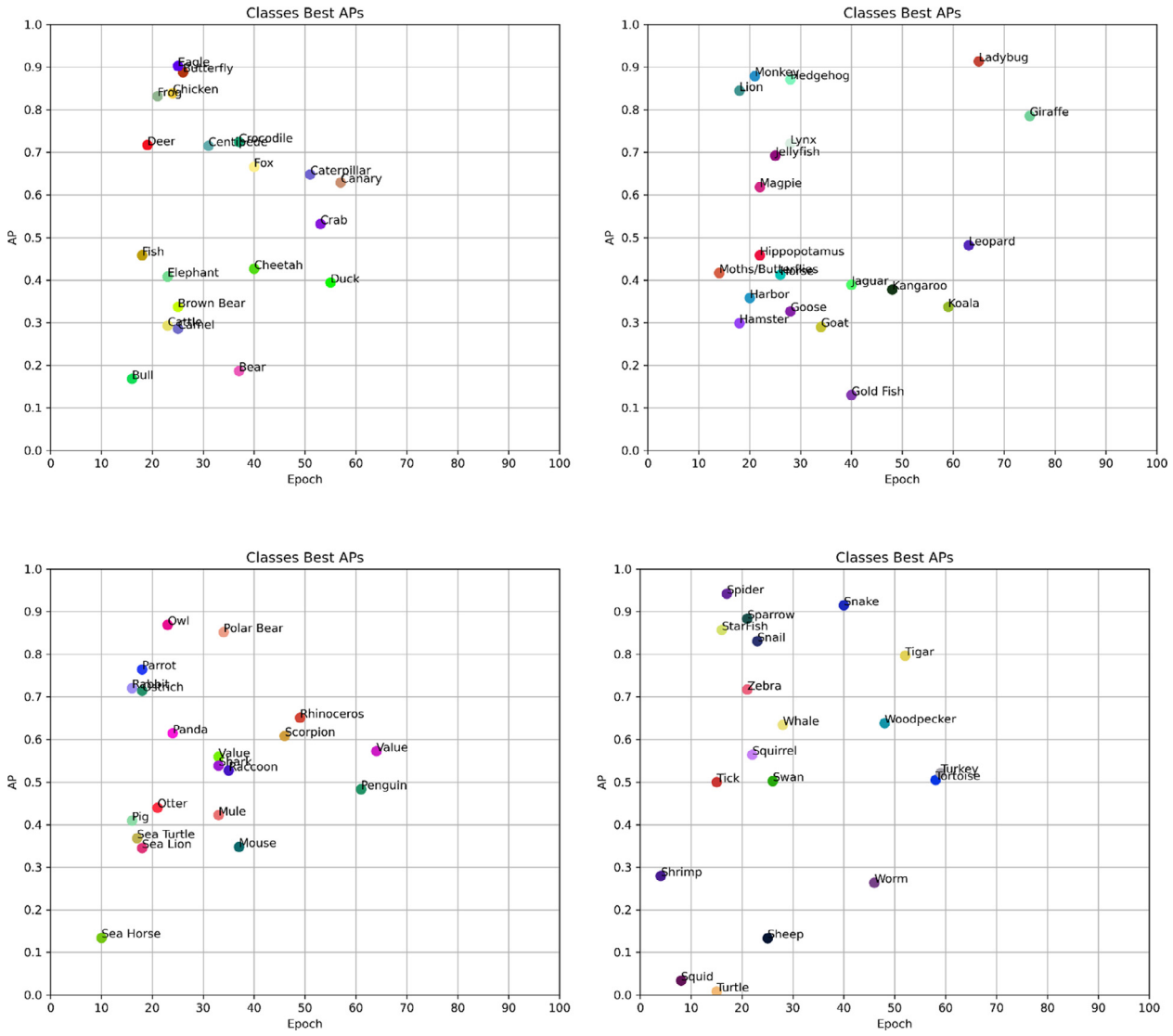


Figure 13. Average Precision Without Attention

Table 5
AP Comparison Result on the African Wildlife Dataset (4 Classes) @Epoch 50

	Cascaded R-CNNR-50-FPN	Faster R-CNNR-50-FPN	RetinaNetR-50-FPN	YOLOv3	OursNo Attention	OursWith Attention
Rhino	0.85	0.88	0.87	0.92	0.90	0.92
Zebra	0.79	0.79	0.83	0.88	0.86	0.88
Buffalo	0.72	0.69	0.79	0.84	0.84	0.86
Elephant	0.70	0.71	0.74	0.81	0.80	0.82
maP	**0.76**	**0.77**	**0.81**	**0.86**	**0.85**	**0.87**

Acknowledgment

This research was supported by the National Natural Science Foundation of China (NSFC) under the project “Development of fetal heart-oriented heart sound echocardiography multimodal auxiliary diagnostic equipment” (62027827).

Reference

[1] T Lin, P Goyal, RB Girshick, K He, P Dollaŕ, Focal loss for dense object detection, in: 2017 IEEE International Conference on Computer Vision (ICCV) 2017, Venice, Italy, 22–29 Oct 2017, 2017, pp. 2999–3007, doi:10.1109/iccv.2017.324. doi:.

[2] W Liu, D Anguelov, D Erhan, C Szegedy, SE Reed, C Fu, AC Berg, SSD: single shot MultiBox detector, in: 14th European Conference Computer vision (ECCV) 2016, Amsterdam, The Netherlands, 11–14 Oct 2016, Proceedings, Part I, 2016, pp. 21–37, doi:10.1007/978-3-319-46448-0_2.

[3] J Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR) 2017, Honolulu, HI, USA, 21–26 July 2017, 2017, pp. 6517–6525, doi:10.1109/cvpr.2017.690.

[4] Y Sakai, H Lu, J.K Tan, H Kim, Recognition of surrounding environment from electric wheelchair videos based on modified YOLOv2, Future Generat. Comput. Syst. 92 (2019) 157–161, doi:10.1016/j.future.2018.09.068.

[5] S Ren, K He, R.B Girshick, J Sun, in: Faster R-CNN: towards real-time object detection with region proposal networks, 39, 2015, pp. 1137–1149, doi:10.1109/tpami.2016.2577031.

[6] J Dai, Y Li, K He, J Sun, R-FCN: object detection via region-based fully convolutional networks, in: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, 5–10 Dec 2016, Barcelona, ,Spain, 2016, pp. 379–387.

[7] T Lin, P Dollaŕ, RB Girshick, K He, B Hariharan, SJ Belongie, Feature pyramid networks for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), Honolulu, HI, USA, 21–26 July 2017, 2017, pp. 936–944. doi.org/10.1109/cvpr.2017.106.

[8] S.U. Sharma, D Shah, Design and Development of Animal Detection Algorithm Using Image Processing," A thesis submitted to Gujarat Technological University for the Award of Doctor of Philosophy in Electronics and Communication Engineering, 2017 https://www.gtu.ac.in/uploads/1Thesis.pdf.

[9] A Strandburg-Peshkin, F.H Jensen, Challenges and solutions for studying collective animal behavior in the wild, Philos. Trans. R. Soc. B 373 (1746) (2022) 20170005, doi:10.1098/rstb.2017.0005.

[10] W Liu, Z Wang, N Liu, N Zeng, Y. Liu, F.E. Alsaadi, "A survey of deep neural network architectures and their applications, J. Neurocomput. 234 (2017) 11–26 April 2017, doi:10.1016/j.neucom.2016.12.038.

[11] X Chen, H Ma, J Wan, B. Li, T Xia, Multi-view 3D object detection network for autonomous driving, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1907–1915, doi:10.1109/cvpr.2017.691. 2017.

[12] B Zoph, V Vasudevan, J. Shlens, Q.V. Le, Learning transferable architectures for scalable image recognition, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8697–8710, doi:10.1109/cvpr.2018.00907. 2018.

[13] B Zhou, A Lapedriza, A Khosla, A. Oliva, A. Torralba, Places: a 10 million image database for scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1452–1464, doi:10.1109/tpami.2017.2723009.

[14] M.C. Stoddard, D Osorio, Animal coloration patterns: linking spatial vision to quantitative analysis, Am. Nat. 193 (2) (2019) 164–186 doi:, doi:10.1086/701300.

[15] E Karpestam, S. Merilaita, A. Forsman, Size variability effects on visual detection are influenced by colour pattern and perceived size, Anim. Behav. 143 (2018) 131–138 2018, doi:10.1016/j.anbehav.2018.07.013.

[16] S. Kumar, S.K. Singh, Monitoring of pet animal in smart cities using animal biometrics, Future Generat. Comput. Syst. 83 (2018) 553–563 doi:, doi:10.1016/j.future.2016.12.006.

[17] G.K Verma, P. Gupta, Wild animal detection from highly cluttered images using deep convolutional neural network, Int. J. Comput. Intell. Appl. 17 (04) (2018 Dec 6) 1850021, doi:10.1142/s1469026818500219.

[18] T Hollings, M Burgman, M Van Andel, M Gilbert, T Robinson, A Robinson, How do you find the green sheep? A critical review of the use of remotely sensed imagery to detect and count animals, Methods Ecol. Evol. 9 (4) (2018) 881–892 doi:, doi:10.1111/2041-210x.12973.

[19] 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.

[20] A. Joly, Lifeclef 2017 lab overview: multimedia species identification challenges, in: International Conference of the Cross-Language Evaluation Forum for European Languages, 2017, 2017, pp. 255–274, doi:10.1007/978-3-319-65813-1_24. doi:.

[21] J Deslauriers, M Toth, A Der-Avakian, V.B. Risbrough, Current status of animal models of posttraumatic stress disorder: behavioral and biological phenotypes, and future challenges in improving translation, Biol. Psychiatry 83 (10) (2018) 895–907 2018, doi:10.1016/j.biopsych.2017.11.019.

[22] L.C Chen, Y Zhu, G Papandreou, F. Schroff, H Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818, doi:10.1007/978-3-030-01234-2_49. 2018doi:.

[23] S.R Kheradpisheh, M Ganjtabesh, S.J. Thorpe, T Masquelier, STDP-based spiking deep convolutional neural networks for object recognition, Neural Netw. 99 (2018) 56–67, doi:10.3410/f.732453284.793575159.

[24] P Badre, S Bandiwadekar, P Chandanshive, A Chaudhari, M.S. Jadhav, Automatically identifying animals using deep learning,", Int. J. Recent Innov. Trends Comput. AL SAADI and El Abbadi Iraqi Journal of Science, 2020 61 (4) (February 2019) 194–197 2361-2370 2370 Commun., doi:10.1021/acs.analchem.9b01891.s001.

[25] Norouzzadeh, Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning, in: Proc. Natl. Acad. Sci, 115, 2018, pp. E5716–E5725, doi:10.14445/23488387/ijcse-v8i5p102.

[26] L.C Chen, Y Yang, J Wang, W Xu, A.L Yuille, Attention to scale: scale-aware semantic image segmentation, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016), 2016, doi:10.1109/cvpr.2016.396.

[27] M Jaderberg, K Simonyan, A. Zisserman, Spatial transformer networks, Twenty-ninth Annual Conference on Neural Information Processing Systems. NeurIPS 2021 is a Virtual-only Conference, 2015. "PDCA12-70 datasheet, Mezzovico, Switzerland, 2021 Opto Speed SA.

[28] Fei Wang, Residual attention network for image classification, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, doi:10.1109/cvpr.2017.683.

[29] J Fu, Z Heliang, M. Tao, Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition, Conf. on Computer Vision and Pattern Recognition, 2017, doi:10.1109/cvpr.2017.476.

[30] X Hongtao, F Shancheng, Z Zheng-Jun, Yating Ya, L Yan, Z Yongdong, Convolutional attention networks for scene text recognition, ACM Trans. Multimedia Comput. Commun. Appl. 15 (Issue 1s) (February 2019) 1–17 Article No.: 3ppdoi.org/, doi:10.1145/3231737.

[31] X Chen, K Kundu, Y Zhu, A.G Berneshawi, H Ma, S Fidler, R Urtasun, 3D object proposals for accurate object class detection, in: NIPS, 2015, pp. 424–432. pages.

[32] J Hu, L Shen, G Sun, Squeeze-and-excitation networks, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, Jun. 2017, pp. 7132–7141, doi:10.1109/cvpr.2018.00745.

[33] F Yang, W Choi, Y Lin, Exploit all the layers: fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, doi:10.1109/cvpr.2016.234.

[34] H Li, Z Lin, X Shen, J Brandt, G Hua, A convolutional neural network cascade for face detection, IEEE Conference on Computer Vision and Pattern Recognition CVPR), 2015, doi:10.1109/cvpr.2015.7299170.

[35] S Bell, K Bala, R Girshick, Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, doi:10.1109/cvpr.2016.314.

[36] X Yu, W Jiangping, K Roland, A.J Patrick, W Tianjiang, H Thomas, Automated Identification of animal species in camera trap images, EURASIP J. Image Video Process. 1 (2013) 52.1-52.10, doi:10.1186/1687-5281-2013-52.

[37] H Nguyen, J.M Sarah, D.N Tu, N Thin, F Paul, A Kylie, G.R Euan, P Dinh, Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring, in: 2017 IEEE International Conference on Data Science and Advanced Analytics, Tokyo, Japan, 2017, pp. 40–49, doi:10.1109/dsaa.2017.31.

[38] D Yudin, S Anton, K Andrey, Detection of big animals on images with road scenes using deep learning, in: 2019 International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI), IEEE, 2019, pp. 100–1003, doi:10.1109/ic-aiai48757.2019.00028.

[39] A.G Villa, S Augusto, V Francisco, Towards automatic wild animal monitoring: identification of animal species in camera-trap images using very deep convolutional neural networks, Ecol. Inf. 41 (2017) 24–32, doi:10.1016/j.ecoinf.2017.07.004.

[40] Z Giraldo, H Jhony, S Augusto, G Alexander, D-P Angelica, Automatic recognition of mammal genera on camera-trap images using multi-layer robust

principal component analysis and mixture neural networks, 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), 2017, doi:10.1109/ictai.2017.00020.
[41] R Girshick, J Donahue, T Darrell, J Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Portland, OR, USA, Jun. 2014, pp. 580–587, doi:10.18127/j00338486-202109-11.
[42] K He, X Zhang, S Ren, J Sun, Deep residual learning for image recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi:10.1109/cvpr.2016.90.
[43] X Wang, R Girshick, A Gupta, K He, Non-local neural networks, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, Jun. 2017, pp. 7794–7803, doi:10.1109/cvpr.2018.00813.
[44] S Woo, J Park, J Lee, I Kweon, CBAM: convolutional block attention module,", in: Proc. Eur. Conf. Comput. Vis., Munich, Germany, 2018, pp. 3–19, doi:10.1007/978-3-030-01234-2_1.
[45] K He, G Gkioxari, P Dollar, R Girshick, Mask R-CNN, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Honolulu, HI, USA, Oct. 2017, pp. 2980–2988, doi:10.1109/iccv.2017.322. doi:.
[46] Z Cai, N Vasconcelos, Cascade R-CNN: delving into high-quality object detection, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, Jun. 2018, pp. 6154–6162, doi:10.1109/cvpr.2018.00644.
[47] V Mnih, N Hees, A Graves, K Kavukcuoglu, Recurrent models of visual attention, in: Proc. Annu. Conf. Neural Inf. Process. Syst., Montreal, QC, Canada, 2014, pp. 2204–2212.
[48] G Huang, Z Liu, L vander-Maaten, K.Q Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 2017, pp. 2261–2269, doi:10.1109/cvpr.2017.243. 21–26 July 2017.
[49] C.C Ukwuoma, Z Qin, H Md-Altab, M.C Bernard, O Ariyo, A.C Ijeoma, J.E Chukwuebuka, S.A Hassan, Holistic attention on pooling based cascaded partial decoder for real-time salient object detection, in: 2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), IEEE, 2021, pp. 378–384, doi:10.1109/prai53619.2021.9551094.
[50] J Zhang, X Zhipeng, S Juan, Z Xin, W Jin, A cascaded R-CNN with multiscale attention and imbalanced samples for traffic sign detection, IEEE Access 8 (2020) 29742–29754, doi:10.1109/access.2020.2972338.
[51] C.C Ukwuoma, B.H Md-Belal, S.M Mahmoud, A Faijan, Z Qin, B.S Emmanuel, A Omar, A Fahed, Image inpainting and classification agent training based on reinforcement learning and generative models with attention mechanism, The 33rd International Conference on Microelectronics (ICM) Cairo, Egypt, 19-, 22 December 2021.
[52] C.C Ukwuoma, B Chen, Deep learning review on drivers drowsiness detection, in: 2019 4th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON), IEEE, 2019, pp. 1–5, doi:10.1109/times-icon47539.2019.9024642.
[53] D.D Cham, T.S Nguyen, Q.M Nguyen, T.T Nguyen, T.D Tran, An analysis of shoreline changes using combined multitemporal remote sensing and digital evaluation model, Civil Eng. J. 6 (1) (2020) 1–10, doi:10.28991/cej-2020-03091448.
[54] A. Sarabu, A.K. Santra, Human action recognition in videos using convolution long short-term memory network with spatio-temporal networks, Emerg. Sci. J. 5 (1) (2021) 25–33, doi:10.28991/esj-2021-01254.
[55] S Arhin, M Babin, B.A. Hamdiat, Predicting travel times of bus transit in Washington, DC using artificial neural networks, Civil Eng. J. 6 (11) (2020) 2245–2261, doi:10.28991/cej-2020-03091615.
[56] 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.
[57] R.M. Nowak, in: Walker's Mammals Of The World, 1, 6th Edition, Johns Hopkins University Press, Baltimore, 1999, pp. 1166–1170.
[58] D.A. Warrell, Venomous bites and stings in the tropical world, Med. J. Aust. 159 (11-12) (1993) 773–779 Dec 6-20, doi:10.5694/j.1326-5377.1993.tb141345.x.
[59] Farhadi A Redmon, YOLOv3: an incremental improvement, arXiv:1804.02767 (2018).