Sensing and Imaging (2024) 25:50
https://doi.org/10.1007/s11220-024-00496-z

RESEARCH

Advancing Wild Deer Monitoring Through UAV Thermal Imaging and Modified Faster RCNN: A Case Study in Nepal's Chitwan National Park

Haitao Lyu[1] · Fang Qiu[1] · Li An[2] · Douglas Stow[3] · Rebecca Lewision[3] · Eve Bohnett[4]

Received: 20 August 2023 / Revised: 30 May 2024 / Accepted: 18 July 2024

Abstract

With traditional survey methods such as ground-based counting, camera trapping, and aerial surveys, monitoring wild deer in Nepal's Chitwan National Park is challenging due to the dense tall vegetation that often conceals them. However, the thermal signatures of wild deer contrast sharply against the cooler background, facilitating detection via thermal imaging. This study explores the use of Unmanned Aerial Vehicles (UAVs) equipped with thermal cameras to monitor wild deer. A large volume of images can be captured, where wild animals appear as small objects. Reviewing these images manually is labor-intensive and time-consuming. To address this, we developed an object detection model using modified Faster R-CNN that automatically identifies small deer objects in the thermal images. Instead of VGG 16, the Feature Pyramid Network and Residual Neural Network (ResNet152) were employed to enhance feature extraction from these images, constructing multi-scale feature maps that enrich the feature information for small object detection. Customized anchor boxes were also designed to handle the wide variation in object scale and aspect ratios. To improve species identification accuracy for small Regions of Interest, a multi-scale aggregation method was proposed, which fuses features from multiple feature maps via Multi-scale RoIAlign pooling. The model proposed in this paper was evaluated by the COCO metrics. The experimental results obtained for the detection of deer and other animals in UAV thermal images with the resolution of 640×512, showing mean Average Precision of 92.3% for all objects, 78.9% for small objects, 94.6% for medium objects, and 95.8% for large objects. This research provides a valuable means for detecting small objects in thermal images and contributes significantly to the field of wildlife monitoring.

Keywords UAVs · Faster R-CNN · Small object detection · Thermal images · Wildlife survey

Extended author information available on the last page of the article

Published online: 13 August 2024

1 Introduction

The urgency to conserve wildlife could never been more emphasized, especially in areas of rich biodiversity such as Chitwan National Park of Nepal, designated as a World Heritage site by UNESCO since 1984. The park covers an expanse of roughly 952,63 square kilometers and harbors a rich tapestry of ecosystems ranging from dense forests and marshlands to rippling grasslands. It's a vital stronghold for a variety of species including the majestic Asian elephant, the one-horned rhinoceros, various deer species, and the wild buffalo and so on. In the intricate web of the Park's ecosystem, wild deer stand as pivotal characters. Their role extends beyond mere presence; as primary herbivores, they are instrumental in the regulation of vegetation through their grazing habits, aiding in seed dispersal and plant growth. This in turn sculpts the landscape, affecting the myriad of other plant and animal life within the park. Moreover, as essential prey for predators like tigers and leopards, wild deer are also critical to the balance of predator–prey dynamics that underpin ecological stability [1]. Therefore, it becomes very important to regularly monitor wild deer and estimate the deer population, which can benefit the wildlife management and conservation in the park.

As for now, several survey methods, including ground-based counting, camera trapping, and remote sensing monitoring, have been proposed to monitor wild deer. Each has its advantages and disadvantages. For example, ground-based counting offers the direct observation of wild deer population and estimation through the signs left by them, such as tracks and poop [16] as shown in Fig. 1a. However, Deer are notably vigilant and tend to avoid humans, complicating the task of achieving accurate counting [11]. Camera trapping automates the photo-taking process, which can be used to capture wild animals when they trigger sensors as shown in Fig. 1b. This method does not disturb wild animals or alter their natural behaviors, and can provide continuous data collection over extended periods, regardless of weather or time of day. However, camera traps are fixed installations, and their placement often relies on human judgment or prior knowledge, which may introduce bias. This could lead to the overrepresentation of certain areas or species, or underrepresentation and even the complete omission of some. Remote sensing technologies are also used in wild animal surveys, like using manned aircraft or satellites to conduct wild animal census. [9] utilized a manned helicopter to monitor and estimate the wild deer

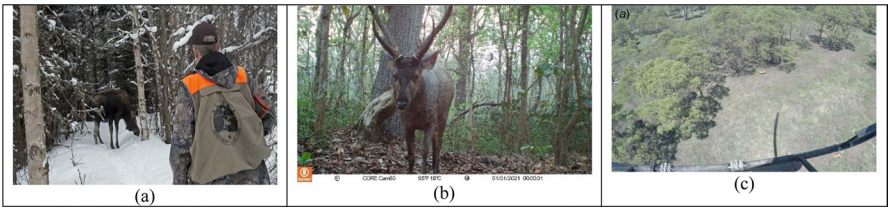


Fig. 1 **a** A wildlife biologist approaches a moose to do ground-based moose survey. ADFG photo. (KTUU) By Grant Robinson; **b** a deer was captured by a camera trapping; **c** a deer was monitored by a helicopter

populations in the Sierra Nevada as shown in Fig. 1c. Compared with ground-based counting and camera trapping, this method can provide wider coverage. However, it is expensive to rent aircrafts and hire skilled pilots. Wild animals are often disturbed by the noise generated by the aircraft.

Among these survey methods, UAV survey gets the best balance, offering convenient data collection and lower cost. At present, using UAVs to monitor wild animals is becoming more and more popular. Additionally, UAVs can be equipped with thermal sensors, enabling the capture of thermal imagery, which is particularly beneficial for monitoring wild animals in the areas of the Chitwan National Park of Nepal covered by tall vegetation and dense tree canopies. An illustrative example is shown in Fig. 2, which presents two images captured simultaneously by a UAV over an area containing 10 wild deer, using two different sensor types. Figure 2a displays a true-color image with a resolution of 8000×6000 pixels, where 3 deer are easily visible in an open space near a tree (marked by green rectangles), 3 deer are partially visible between tree canopies (yellow rectangles), and four deer are obscured by tree canopies (red rectangles) and not visible. Figure 2b showcases a thermal image with a resolution of 640×512 pixels, where all 10 deer are visible, including those under tree canopies, due to their body temperatures being higher than the ambient background.

Our research area is in Chitwan National Park, Nepal. Since the park's landscape is predominantly covered by trees and tall grasses, poses a challenge for true-color sensors due to vegetation occlusion, UAVs mounted thermal cameras were chosen to monitor wild deer in our study. UAVs make it much easier to capture wild animals, leading to a significant increase in the volume of captured thermal images. Consequently, it is more and more impractical to review these images manually to count the number of wild animals. Deep learning, which is one of the cutting-edge techniques, has been wildly used in the detection of wild animals from terrestrial images. For instance, in [21], ResNet50 was used to detect animals from true-color images, and they achieved a classification accuracy of 93.8% on the

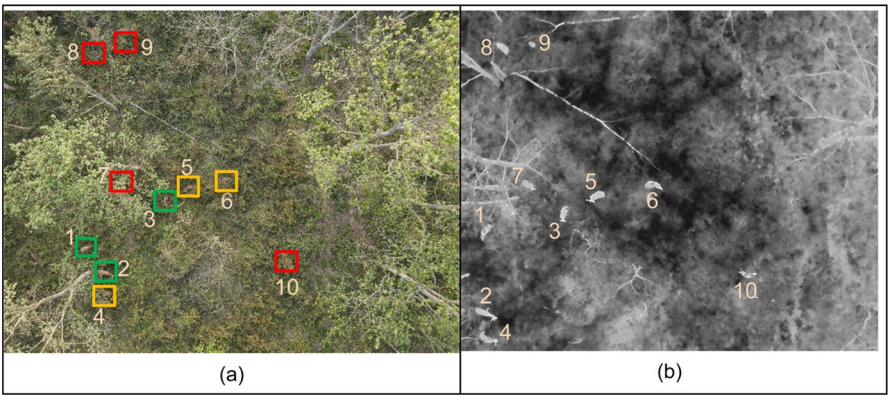


Fig. 2 Four wild deer marked by *red boxes* were not seen on the 8K true-color image in **a** due to canopy cover, but they were visible on the thermal image in **b** due to their higher body temperature

Snapshot Serengeti dataset in Africa. In [3], a pre-trained Faster RCNN based on InceptionResNetV2 network was utilized to detect European mammals from high-resolution images and achieved a detection accuracy of 94%. In [7], YOLOv5 was employed to detect Red deer from images captured by cameras, achieving a detection accuracy of 0.86. These methods can automate the process of wild animal detection from images. Among these studies, the advanced object detection models based on deep learning, like Faster R-CNN and YOLO have gained considerable success, when high-resolution cameras were utilized to capture wild animals in areas that are not covered by tall vegetation. The representation of wild animals in these high-resolution terrestrial images are large and clear as shown in Fig. 3 so that they can provide enough feature information for these models to localize their positions and identify their species. However, these models have many limitations in the detection of wild animals from thermal images captured by UAVs as shown in Fig. 2b. These limitations can be summarized as three research problems.

- **Limited feature information**: compared with deer objects in high-resolution images, the deer objects in thermal images is small and unclear due to their limited resolution, making it difficult to provide enough information for object detection models to localize their positions.
- **Large span of object scale variation**: the terrain of our research areas in the Chitwan National Park of Nepal is uneven with ups and downs. UAVs often need to fly at different altitudes to avoid the top tips of trees. The variability in different flight height results in the thermal imagery with scales of observed wild animals varying significantly. For example, the smallest deer object in a thermal image in our study is about 100 pixels, and the biggest deer object is beyond 6400 pixels. The difference is about 64 times, which shows a big challenge for current models to detect them.

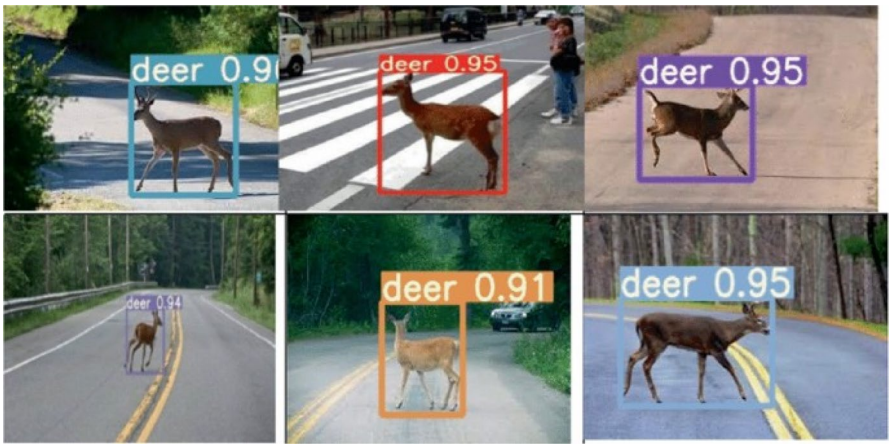


Fig. 3 The representation of deer in the true-color images captured by high-resolution cameras

- **Species Identification based on RoIs**: due to the small size of wild animal objects in thermal images, the RoIs associated with these objects are also small, making it challenging to extract sufficient features for species identification.

In this paper, an object detection model based on Faster R-CNN was constructed to automatically detect wild deer from the thermal images captured by UAVs. Due to the challenge that small size of wild animal objects in thermal images cannot provide enough feature information for localization and identification, Feature Pyramid Network (FPN) and Residual Neural Network (ResNet152) instead of the commonly used VGG16, were used to extract feature information from thermal images to construct multi-scale feature maps. Each feature map can contribute to the quantity of feature information, and by adding the number of feature maps, the problem of limited feature information for small object detection can be mitigated. According to the large span of object scale variation, customized anchor boxes were used to enable our model to detect the objects across a wide range of scales and aspect ratios. To cope with the difficulty of species identification with small Regions of Interest (RoIs), a multi-scale aggregation method was proposed to fuse the feature of RoIs from multiple feature maps extracted by Multi-scale RoIAlign pooling, which can improve the precision of deer species identification. The rest of the paper is organized as follows. Section 2 provides an overview of the history of using information technology to automate the task of detecting animals from images. In Sect. 3, we describe the details about how to modify the architecture of ResNet152 and FPN to construct an object detection model based on Faster R-CNN, making it more suitable for small object detection from low-resolution thermal images captured by UAVs. Section 4 discusses the application of the model to a thermal image dataset collected from the Chitwan National Park, presenting experimental results and analysis. Finally, the paper concludes with a summary of the key findings and contributions.

2 Related Work

Over the past decade, using information technology to detect wild animals from images has remained a hot research spot, especially in the field of ecologic informatics. Many methods have been proposed, which have led to significant advancements. These methods can be broadly categorized into two types: pixel-based and region-based.

2.1 Pixel-Based Method

Pixel-based methods involve a series of steps to process the images, including image pre-preprocessing, image segmentation, object detection, feature extraction, feature selection, and object classification. The performance of each stage significantly influences the final classification accuracy. In addition, expert knowledge plays a vital role in these methods. Experts are required to identify meaningful features for

the desired classification, representing unique characteristics of wild animals and their habitats. These features are then utilized by some algorithms designed to calculate the differences between animal objects and their surroundings through statistical analysis. In the early stages, threshold setting is a simple and widely used approach to differentiate animals from images. The idea behind this method is to apply a threshold value to a specific image feature, such as color, intensity, or texture, and then consider the regions where the pixels surpass the threshold [4, 36]. For example, in [13], a pixel-based classification method was proposed to classify pixels based on their spectral characteristics and compare them to predefined threshold to find the regions including animals. The spectral thresholds based on light and dark pixels were used in [4] to identify white birds from RGB images that had dark backgrounds. Similarly, based on the values of pixels in infrared imagery, [8] utilized the threshold of temperature to distinguish hot rabbits from a cold background. Moreover, blue/green thresholds have proven useful in classification between marine mammals and water. This approach has been applied in multiple studies to effectively differentiate marine mammals from their aquatic environment, including [24, 29], and [28]. Pixel-based methods are easy and can be quite effective when the target animals in images have distinct features. However, in complex environments, wild animals are often camouflaged or their features blend into the background, making it difficult for pixel-based methods to detect them. Therefore, more advanced pixel-based methods based on machine learning were proposed [6]. For instance, Oriented Gradients (HOG) and Haar-like features, along with classifiers like Support Vector Machines (SVMs) were used to detect animals from the images captured from complex environments[25]. Torney et al. [32] introduced a method that combined rotation-invariant object descriptors with machine learning algorithms to detect wildebeests from aerial images. Object-based Image Analysis (OBIA) approach involves DBSCAN cluster algorithm to group similar pixels into contiguous objects, which was used in [5] to count birds in large volumes of aerial imagery. In [27], a supervised pixel-based image classification model demonstrated high accuracy in counting Lesser Black-backed Gulls and hippopotami in homogeneous environments with no obstructing vegetation. However, these approaches often exhibit a lot of limitations in complex environments where animals blend with their surroundings.

2.2 Region-Based Method

Region-based methods utilize deep learning techniques, especially Convolution Neural Networks (CNNs), to process and analyze image data at a more contextual level than pixel-based approaches. CNNs can directly extract hierarchical features from images, which are essential for recognizing and differentiating objects within diverse and cluttered backgrounds. Some advanced object detection models based on CNNs have been proposed, such as Faster R-CNN [26], SSD [19], RetinaNet [18], and YOLO [2]. These models have been used in the detection of wild animals from images. For example, a deep neural network using ResNet50 as feature extractor was used to identify wild animals in the Snapshot Serengeti's true-color images obtained via camera traps in Africa, achieving a detection accuracy of 93.8% in [21].

A model combining FPN and ResNet50 was employed to detect elephants, giraffes, and zebras from high-resolution terrestrial images in Kenya's Tsavo National Park with the detection accuracies of 95% for elephants, 91% for giraffes, and 90% for zebras in [10]. A pre-trained Faster RCNN based on InceptionResNetV2 detected European mammals in camera trap images with a 94% accuracy in [3]. [23] used Faster R-CNN to identify kiang in high-resolution drone images, achieving an overall precision of about 90%. In [7], YOLOv5 was employed to detect Red deer from images captured by mobile cameras, achieving a detection accuracy of 0.86. [33] utilized ResNet50 as backbone to construct Faster R-CNN to detect deer and boars from the images captured by camera trapping, obtaining a detection accuracy of 0.88.

2.3 Summary

At present, the majority of research is concentrated on detecting wild animals in true-color (RGB) images captured by using high-resolution cameras, and some models demonstrated excellent performance in the detection of wild animals. However, the research area of detecting small objects in thermal imagery is still relatively underexplored. For example, compared with the animal objects in high-resolution true-color images shown in Fig. 3, the size of wild animal objects in thermal images captured by UAVs is much smaller as shown in Fig. 2b, which means it may not provide enough feature information for the object detection models using deep learning to localize their positions and identify their species. In addition, when UAVs are used to monitor wild animals in the areas covered by tall vegetation, they have to fly at different altitudes to dodge the tips of tall trees. As a result, animal objects in thermal images may appear at different distances from the cameras, leading to significant variations in object size within a thermal image dataset. For instance, the animal objects in the UAV thermal imagery used in this paper range in size from 10×10 pixels to 80×80 pixels, with the largest object being nearly 64 times the size of the smallest. It is difficult for the ready-to-use object detection models designed for true-color imagery to recognize them. Therefore, as the use of UAVs for monitoring wild animals becomes increasingly popular, it is crucial to conduct more research into methods for automating the detection of small animal objects in the thermal images with limited resolution.

3 Methodology

3.1 Overview

In this paper, we apply the Faster R-CNN architecture to detect and classify wild animals in thermal images taken by UAVs. Faster R-CNN, originally proposed by [26], is not merely an object detection model but a comprehensive two-stage framework for detecting objects. This framework includes several key modules: Feature Extraction, Region Proposal Network (RPN), Anchor Boxes, Bounding Box Regression,

and Region of Interest (RoI) Classification. To some extent, the process of object detection using deep learning emulates the process the humans use their eyes to find objects and then use their brains to identify the types of the objects.

Initially, the Feature Extraction module processes an image using a grid to create a feature map, as shown in Fig. 4. Each unit of this grid correlates to a specific area of the input image, even though after transformations and downsampling being performed on them by the convolutional neural networks (CNNs). Using this feature map, the RPN predicts potential object locations, where anchor boxes are placed. These boxes, functioning as the model's eyes and essential for the model's identification capability, are then refined through Bounding Box Regression to make them fit the objects optimally. Finally, the identified RoIs are passed into the RoIs Classification module to determine the species of the detected animals. As shown by the process of the object detection model in Fig. 4, feature maps play a fundamental role on how a model can perceive, process, and interpret an image. Feature maps can be produced at different stages of CNNs through convolutional operations through the filters or kernels applied across the input image or the output from previous layers. Each filter is designed to detect specific types of features at different levels of abstraction, such as edges, textures, colors, or more complex shapes or patterns, which are important for object classification. In addition, feature maps also maintain the spatial hierarchy of the input image, which is essential for the detection of the location of objects within an image.

Compared with high-resolution true-color images, thermal images have lower resolution and inherently lack details, which makes it difficult for CNNs in Faster R-CNN to extract meaningful features to detect small objects within them. To address this limitation, we have implemented a series of modifications to the standard architecture of Faster R-CNN, enabling it to effectively detect small objects in UAV thermal images. The detailed structure of the detection model based on Faster R-CNN proposed in this paper is shown in Fig. 5. Subsequent subsections will further elaborate on these modifications.

3.2 Feature Map Extractor Based on FPN and ResNet152

Classical feature extraction networks used in object detection typically include: LeNet [17], which was the first convolutional neural networks applied to solve practical problems; AlexNet [15] which introduced dropout to prevent overfitting and proposed the ReLU activation function; VGG network [30], which utilized modularization in its design and replace the larger convolutional filters with multiple 3×3 filters; ResNet [12], which introduced a residual network structure to solve the issue of gradient explosion and disappearance as the number of CNN layers increases, allowing for very deep network layers; GoogleNet [31], which introduced the inception module and adopted multiple branches and convolution kernels; and ResNeXt [35], which integrated the concepts of Inception and ResNet.

GoogleNet and ResNext are usually used for object detection in high-resolution images, significantly reducing the number of parameters by executing convolutional operations across multiple branches, thereby enhancing model efficiency.

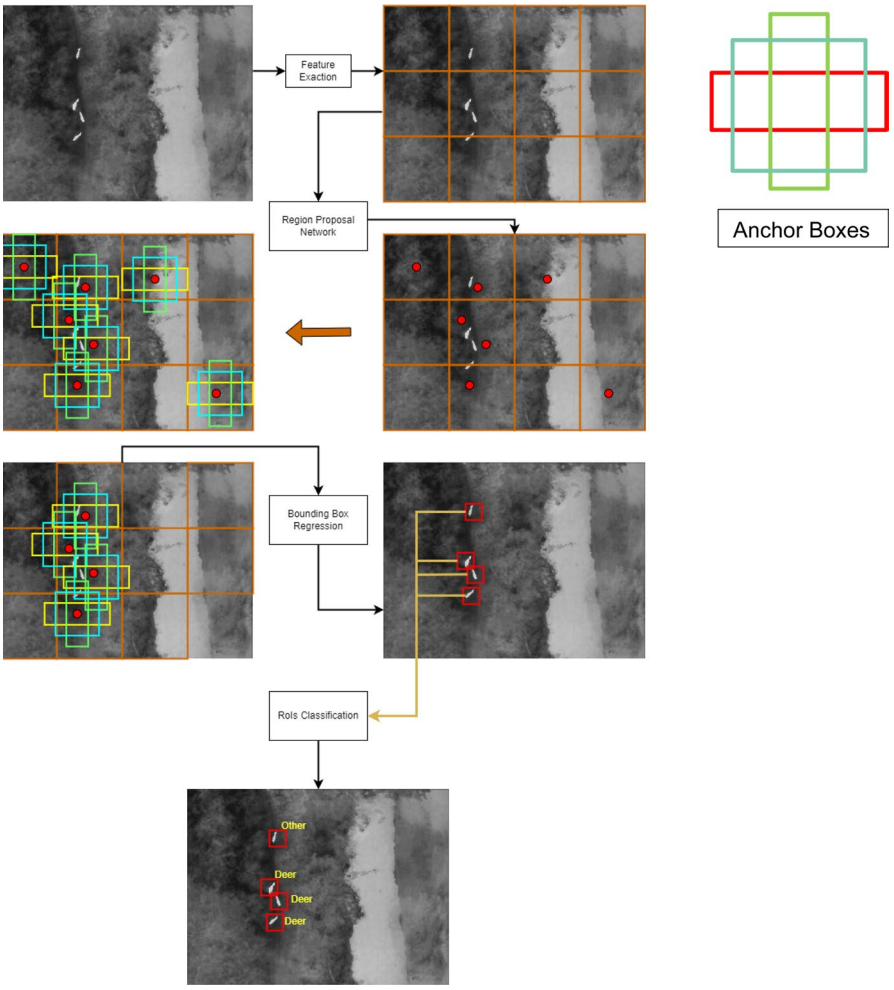


Fig. 4 The framework of Two-stage Object Detection

However, these architectures may not be as effective for low-resolution images due to the inherently limited feature information available. Branching operations in such contexts can further dilute the feature details on each branch, leading to suboptimal detection outcomes. Conversely, VGG and ResNet are preferred for object detection in low-resolution images. VGG16 was used in the original Faster R-CNN in [26] to extract feature information from input images, but with its sequential architecture shown in Fig. 6a, may cause the gradients to diminish as they backpropagate through each deeper layer, making it difficult to detect small objects from low-resolution images. In this study, the resolution of a thermal image is 512×640, and the average size of objects in it is about 25×25 pixels. With the downsampling process of four CNN layers in VGG, the spatial dimension of the object in the feature map is

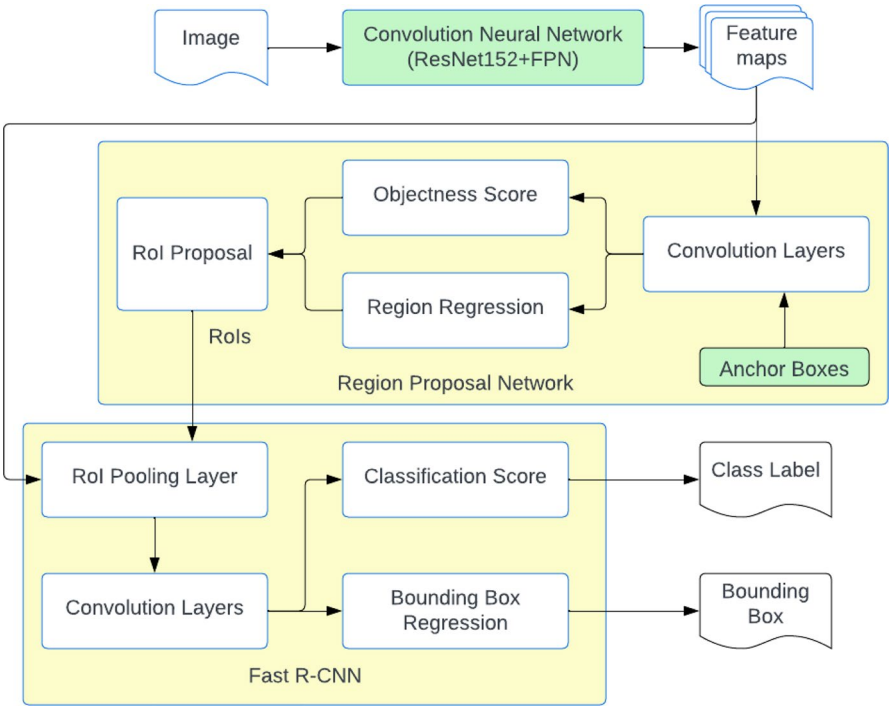


Fig. 5 The Structure of Modified Faster R-CNN in this paper

condensed to be less than 2×2 pixels. With the CNN layers going deeper, the feature of small objects disappears, and only background-related information remains in the feature map, leading to smaller and smaller gradient values. On the contrary, ResNet is made up of residual blocks that include shortcut connections as shown in Fig. 6b. These connections allow gradients to flow directly through the network during training, mitigating the vanishing gradient problem. Therefore, ResNets can be trained with much deeper layers than VGG networks. According to the number of CNN layers, the family of ResNets consists of ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152. In general, reception fields of a feature map are crucial for the classification of small objects. In deep CNNs, layers are stacked, and each layer's receptive field builds on the previous layers, thus, deeper layers in a CNN tend to have larger receptive fields. Given the small size of wild animal objects in our thermal images and the importance of receptive fields in identifying their speices, the ResNet152 architecture is chosen as the foundation for the feature extractor eventually.

As shown in Fig. 7, with the CNN layers deepening, the receptive field size increases due to accumulated effects of multiple convolution layers. However, the spatial dimensions of the feature maps typically decrease. This reduction has negative effects on localizing the positions of small objects but has positive effects on the identification of speicies. At the same time, the Field of View of a feature map is calculated by the resolution of an original image divided by its feature map's

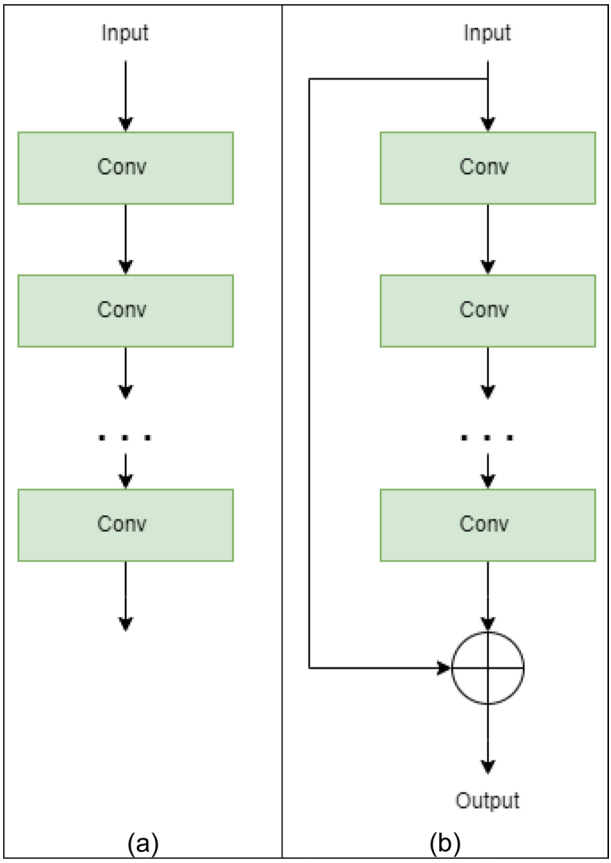


Fig. 6 **a** Sequential Connection; **b** Shortcut Connection

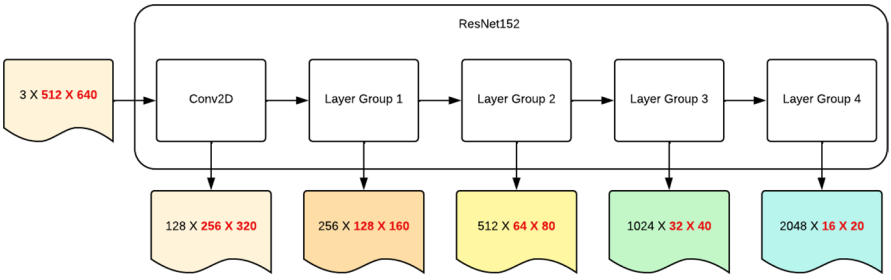


Fig. 7 The structure of ResNet152 Feature Extractor; The red numbers represent the resolution of a feature map. For example, the output of Layer Group 1 is $256 \times 128 \times 160$, which means its resolution is 128×160 and the number of channels is 256

resolution. Therefore, early layers with small field of view might only see small, local parts of the input, while deeper layers with big field of view may encompass a broad view of the input. Different feature maps can contribute to the detection of objects with different size. For the task of small object detection, a popular method is to select the high-resolution feature map for the Faster R-CNN model. However, the animal objects in our study are not only small, but also vary significantly in scales and aspect ratios. Only using one feature map cannot work well and thus the five feature maps shown in Fig. 7 are all chosen for the Faster R-CNN to perform the detection task. In addition, FPN is used to fuse the information from the five feature maps, to allow the early feature maps to complement the deeper feature maps. Therefore, the integration of FPN and ResNet152 is used in this paper as shown in Fig. 8. Specifically, the feature maps with different scales are generated by the different layers in ResNet152. Then FPN is used to combine low-resolution, semantically strong features with high-resolution, semantically weak features via a top-down pathway and lateral connections across the feature maps.

The five feature maps generated by ResNet152 shown in Fig. 7 are denoted as $\{C_1, C_2, C_3, C_4, C_5\}$. In the FPN built on top of ResNet152 (Fig. 8), the features from the five feature maps are merged along top-down pathway. Specifically, the feature map $C_i (i = 1,2,3,4)$ undergoes an upsampling process, increasing its resolution by a factor of 2. The upsampled output is then combined with the corresponding bottom-up feature map $C_j (j = 2,3,4,5)$ using element-wise addition. This merging operation allows for the integration of high-dimension features from the upsampled feature map C_i with the features from the feature map C_j. Additionally, to mitigate the potential aliasing artifacts resulting from the merging operations, we applied a

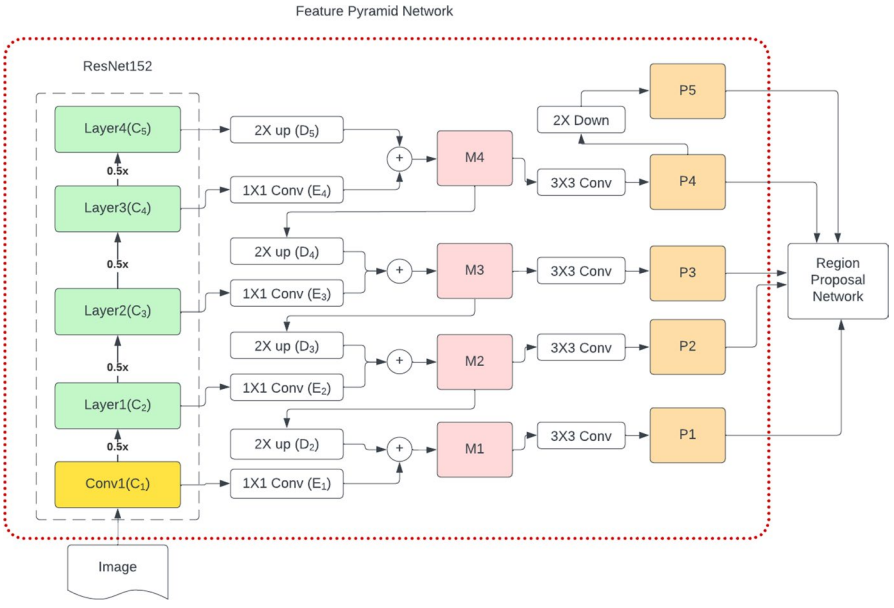


Fig. 8 The structure of FPN based on ResNet152

convolutional operation with the kernel size of 3×3 to the feature maps merged by C_i and C_j, which outputs the final feature maps used by Faster R-CNN model. For instance, according to Fig. 7, the feature map C_5, which is the output of Layer 4, is upsampled to be twice of its original size, as indicated by D_5. Then, a 1×1 convolutional layer is applied to the output of Layer 3 to change its channel dimensions, and the output, denoted by E_4, has the same number of channels as C_5. Through element-wise addition, M_4 is generated, satisfying the equation $M_4 = D_5 + E_4$. Subsequently, a 3×3 convolutional operation is applied to create the feature map P_4. M_4 is then upsampled by a factor of 2 to become D_4. The output of Layer 2 is processed through a 1×1 convolutional layer to produce E_3. By element-wise addition, M_3 is generated, satisfying the equation $M_3 = D_4 + E_3$. Similarly, M_2 and M_1 are created, respectively fulfilling the equations $M_2 = D_3 + E_2$ and $M_1 = D_2 + E_1$. Subsequently, 3×3 convolutional operations are applied to M_1,M_2,M_3, and M_4 to generate four feature maps, denoted as P_1,P_2,P_3, and P_4. Finally, P_5 is produced by downsamplingP_4. The five feature maps, represented as $\{P_1, P_2, P_3, P_4, P_5\}$, are then inputted into Faster R-CNN for the detection of objects.

3.3 Customized Anchor Boxes for Small Object Detection

The concept of anchor boxes, introduced by [26], involves using a set of predefined bounding boxes of various scales and aspect ratios distributed methodically across a feature map. Each anchor is positioned strategically throughout the feature map and then mapped back onto the original image to help align these anchor boxes with the actual ground-truth bounding boxes that encompass the targets. This setup enables an object detection model to adaptively identify objects of different sizes and shapes by using anchor boxes as reference templates that provide vital spatial context to enhance the detection process. At the same time, Intersection over Union (IoU) is a measure used to measure the overlap between two bounding boxes. It is defined as the area of overlap between the predicted bounding box and the ground-truth bounding box divided by the area of union of these two boxes:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$$

In the original Faster R-CNN model, hand-picked anchor boxes were used and each anchor on the feature map is generated in one of nine configurations, combining three scales $\left(128^2, 256^2, 512^2\right)$ and three aspect ratios (0.5,1, 2). This rigid configuration may not align well with the actual sizes and shapes of small animal objects in thermal images, particularly when these objects vary widely due to distance, angle, or environmental factors. During the training of Faster R-CNN, IoU is used to determine which anchor boxes best correspond to a ground-truth object. Anchor boxes with an IoU exceeding a certain threshold (commonly set at 0.7) with a ground-truth box are considered positive examples (true positives), while those with an IoU below a lower threshold (often around 0.3) are treated as negatives (true negatives). As illustrated in Fig. 9, the Intersection over Union (IoU) values between the green anchor boxes and the red ground-truth box are all below 0.7. This

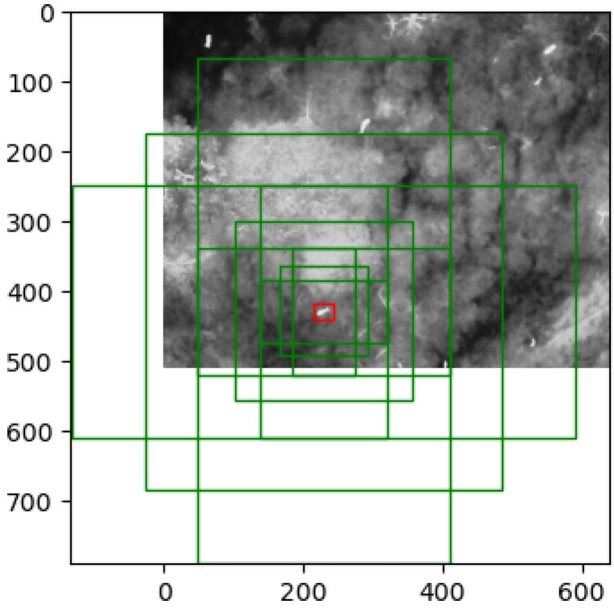


Fig. 9 An example of using anchors to detect deer. The *red rectangle* is a ground-truth bounding box, and the *green rectangles* are the anchors generated by three scales $\left(128^2, 256^2, 512^2\right)$ and three aspect ratios (0.5,1, 2)

indicates that none of the green anchor boxes sufficiently match the ground-truth object, resulting in the failure to detect this animal.

To address this problem, Clustering of Bounding Boxes is used in this paper instead of hand-picked anchor boxes. This method was first used in YOLOv2 and achieved good performance. The main idea is to employ K-means clustering on the training set bounding boxes to determine the dimensions of the anchor boxes. This clustering approach aims to automatically find the most representative anchor box shapes and sizes based on the actual data distribution, which helps in improving the model's accuracy and efficiency. K-means clustering iteratively updates the centroids of clusters by minimizing the total within cluster variance based on the chosen distance metric. The convergence of this process depends heavily on the distance that defines how close the anchor boxs are from each other. Traditionally, the distance between two anchor boxes based on their width and height is calculated to measure their similarity. When applied to anchor boxes, the formula is defined as $\sqrt{\left(\omega_2 - \omega_1\right)^2 + \left(h_2 - h_1\right)^2}$, where ω and h represents the width and height of the anchor boxes respectively. This distance quantifies the absolute geometric difference between the size of two anchor boxes. It is simple and fast to compute but does not consider the position or overlap between boxes. Therefore, IoU is used to measure the similarity between two anchor boxes, and the distance between two boxes in our study is then defined as $1 - IoU$. Unlike the traditional distance, which purely considers

Table 1 Average width and height of nine bounding box clusters labeled in the thermal image dataset

Cluster	1	2	3	4	5	6	7	8	9
Width	12	11	17	15	24	19	19	28	34
Height	11	15	14	19	13	19	26	23	44

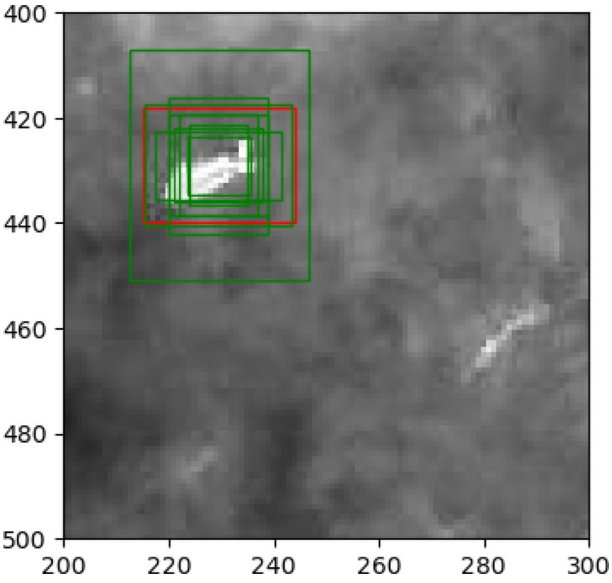


Fig. 10 An Example of Using Customized Anchors to Detect Deer

the dimensions (width and height), IoU accounts for how well two anchor boxes overlap. The distance based on IoU thus reflects not just the size but also how similar the positioning of the two boxes is. This is more suitable for detecting small objects, as the primary concern is whether any part of the anchor captures the object, rather than the exact fit of the dimensions.

We utilized the K-means algorithm based on the IoU distance metric to divide these bounding boxes into nine clusters. The average width and height of these clusters are shown in Table 1, which can be used to generate 9 customized anchor boxes. Figure 10 shows an example of using customized anchor boxes to detect a deer object. The red rectangle is a ground-truth bounding box, and the green rectangles are the anchors generated based on the 9 average cluster width and height listed in Table 1. Among the nine green anchor boxes, at least one has an IoU value greater than 0.7 with the red ground-truth bounding box, ensuring higher detection precision for deer objects in thermal images.

3.4 Fusion of Multi-scale RoI Align for Species Identification

In our paper, a FPN based ResNet152 is used to process the input image through multiple convolutional layers. This creates several feature maps, each representing the input image at different scales and levels of abstraction. These feature maps contain comprehensive spatial and semantic feature information about the input image. Despite transformations, each point in a feature map corresponds to a specific region in the input image, known as its receptive field. As the network goes deeper, the receptive field associated with each point in the feature map covers a larger area of the input image.

As shown in Fig. 5, Regions of Interest (RoIs) are initially identified in the input image through the Region Proposal Network (RPN). These RoIs can be mapped back to corresponding locations on the feature maps. Thus, specific sub-sections of the feature map that correspond to the RoIs can be clipped or cropped, ensuring that only relevant features are obtained for species identification of animal objects within the RoIs. In the original Faster R-CNN and its modified versions proposed in other studies, only one sub-section of the feature map is chosen based on the size of the RoI. Typically, a larger RoI is assigned to a smaller-scale feature map, while a smaller RoI is assigned to a larger-scale feature map. However, this approach may not provide sufficient feature information to identify species of small objects in low-resolution thermal images.

To address this issue, we propose a method called fusion of multi-scale RoI align. As depicted in Fig. 11, an RoI in a thermal image is mapped to four feature maps with different scales. Four sub-sections from these feature maps are then obtained. Using the RoI Align layer, these irregularly sized RoI feature map clips are transformed to a fixed size. Finally, the four fixed-sized RoI feature map clips are fused into one feature map clip, which is used for subsequent species identification. This fusion method allows the output RoI feature map clip to learn feature information from multiple feature map clips, enhancing the precision of species identification of wild animals in thermal images.

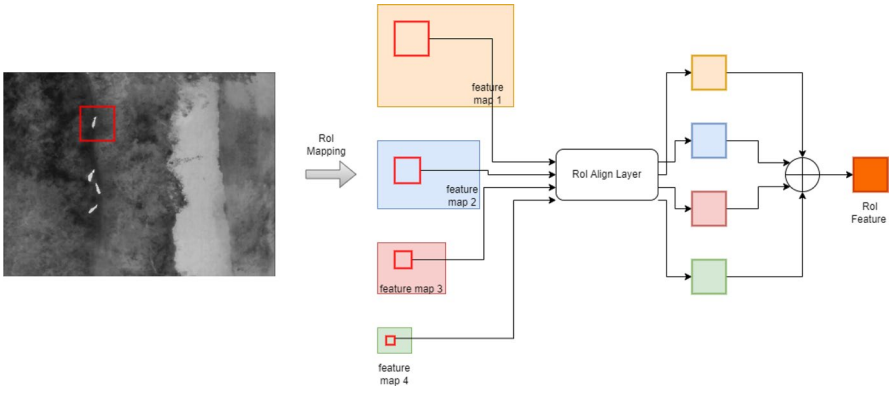


Fig. 11 Flow of Fusion of Multi-scale RoI Align

4 Experiment

4.1 Data Allocation

A total of 22,478 thermal images were captured in Chitwan National Park, Nepal, with all images standardized to a 'White-Hot' palette. To create the training dataset, two scientists from the Center for Complex Human–Environment Systems at San Diego State University, along with two students from the University of Texas at Dallas, participated in the labeling task. Initially, the scientists independently filtered the images to retain only those with deer. Following this, the students used 'ImageLab,' a free online image annotation tool, to label the deer objects in the images identified by the scientists. Any other animals present were annotated as "Other." The scientists then meticulously reviewed all annotations to validate them either as ground truth or to discard them. This thorough verification process significantly reduced subjective biases and ensured the annotations' accuracy and reliability. As a result, a dataset consisting of 5,651 thermal images with 27,403 wild animal instance annotations was constructed.

The number of wild animal instances in each thermal image ranges from 1 to 84. As shown in Fig. 12, there are 62 wild animals in the image, labeled by red boxes. If the dataset is randomly divided into training, validation, and testing subsets, the distribution of wild animal numbers within each subset may differ. This discrepancy can impact the final validation and estimation of the model. To address this issue, the dataset is initially divided into separate subsets based on the number of wild animals in each image. Within each subset, the images are further partitioned into three sections using a ratio of 70%:15%:15% (14:3:3). The detailed distribution is shown in Table 2. Subsequently, the images from these sections are combined to form three distinct datasets, which are allocated for training, validation, and testing, respectively. By employing this approach, we ensure that each dataset maintains a representative distribution of wild animal instances, enhancing the accuracy and reliability of the model trained by them.

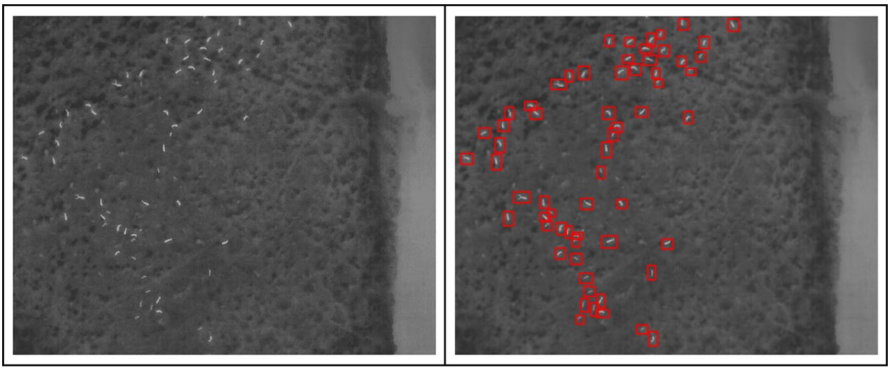


Fig. 12 An example of wild animal annotation. There are 62 wild animals in the left image. The 62 deer are annotated by red boxes

Table 2 Distribution of datasets for training, validation, and testing

	Deer	Other	Sum
Train (70%)	9359	9652	19,011
Test (15%)	1986	2150	4136
Validation (15%)	2164	2092	4256

4.2 Detection Model Evaluation

To ensure a comprehensive assessment of the model's performance, this paper utilized the COCO detection evaluation matrix [22]. All objects were classified into three categories based on their areas: large, medium, and small. The detailed distribution of bounding box areas within our dataset is illustrated in Fig. 13. Small objects have bounding boxes with areas ranging from 0 to 200 square pixels, medium objects range from 200 to 400 square pixels, and large objects have areas exceeding 400 square pixels. Figure 14 shows three different thermal images to illustrate the shapes of small, medium, and large deer objects.

Totally, all animal objects detected from thermal images are classified into two classes: Deer and Other Animal. Average Precision (AP) and Average Recall (AR) are computed across the two categories using IoU thresholds that range from 0.5 to 0.95 with a step size of 0.05. This range accounts for varying levels of overlap between the predicted and ground truth bounding boxes. AP measures the model's accuracy in making positive predictions, with a focus on minimizing false positives. It assesses how well the model's positive predictions align with the ground truth. Conversely, AR measures the model's ability to correctly identify all positive instances, emphasizing the reduction of false negatives. It evaluates how comprehensively the model captures all relevant objects. Together, AP and AR provide insights into the trade-offs between precision and recall, offering a method to assess the overall effectiveness of the model.

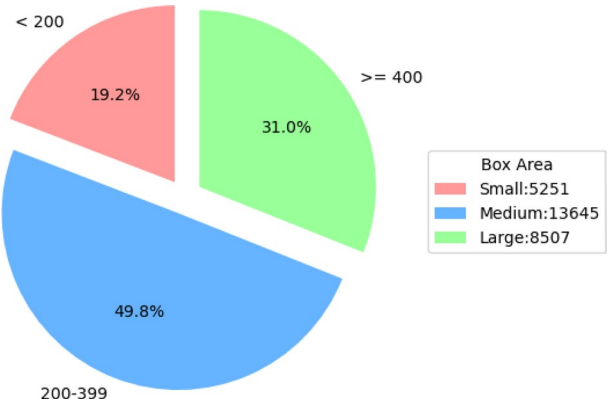


Fig. 13 Distribution of Bounding Box Areas in Our Dataset

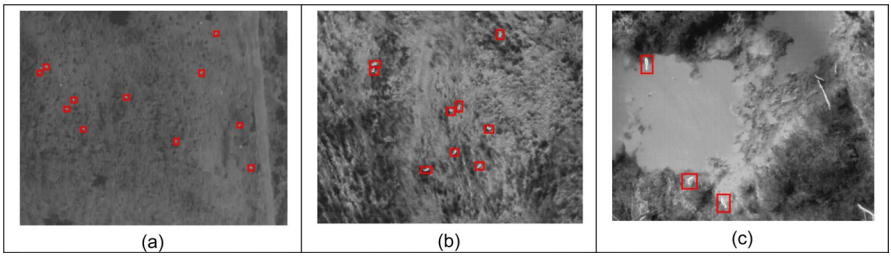


Fig. 14 **a** An example of small objects; **b** an example of medium objects; **c** an example of large objects

4.3 Model Validation

For the tasks of object detection, the family of residual neural networks (ResNets) includes ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152 are very popular as backbones to extract feature information from images because of their structure of Shortcut Connection shown in Fig. 6b. In [20], ResNet152 was proved to be very suitable for the detection of small objects from thermal images. Therefore, in this study, three main tactics are adopted to modify the original Faster R-CNN model. Firstly, VGG16 is replaced by FPN + ResNet152. The original object detection model using only a singular feature map is updated by a new model that can utilize multi-scale feature maps to detect small deer objects in thermal images. Secondly, customized anchor boxes defined in Table 1 is used to replace large-scale anchor boxes $\left(128^2, 256^2, 512^2\right)$, which makes the model more effective in small object detection. Thirdly, the resolution of the RoI pooling layer output is changed from 7×7 to 8×10, keeping the same aspect ratio as the input thermal images with a resolution of 512×640. In addition, the feature information of RoIs, obtained by fusing the outputs from multiple RoI Pooling layers with different scales, is used for species identification in place of only using one output of a RoI Pooling layer.

Based on the structure of Faster R-CNN, four models, denoted as M1, M2, M3, and M4, were constructed with different strategies of backbone networks, anchor boxes, and RoI Pooling. In addition, we also set up a model based on the latest YOLOv8, as M5. The backbone of M5 employs the structure of Spatial Pyramid Pooling—Fast (SPPF) layer [14]. Table 3 provides details and specifications for the five models. Table 4 shows the validation results of the five models based on the COCO evaluation matrix. Based on the scores of Average Precision with a threshold of $IoU \geq 0.5$, the Faster R-CNN model detection performance, ranked from highest to lowest, is as follows: M1, M4, M3, and M2. The detection performance of M1, M3, and M4 is obviously better than M2. As shown in Fig. 15, a thermal image contains four two large deer objects, one medium deer object, and one small deer object. Based on this image, we test the converge speed of the four models. When trained for 100 epochs, Fig. 16 illustrates the detection results of the four models. According to Fig. 16, M2 failed to detect all objects, M3 successfully detected all large objects but struggled to detect medium and small objects, while both M1 and M4 exhibited the ability to detect all objects. However, it's worth noting that M4's

Table 3 The main structures and key characteristics of the four models in the subsequent experiments

Name	Backbone	RoI pooling	Basic anchors
M1	RestNet152 + FPN	8 × 10; Multi-scale RoIs fusion	Customized anchor boxes
M2	VGG16	7 X 7; Single RoI	Large anchor boxes
M3	RestNet152 + FPN	8 × 10; Multi-scale RoIs fusion	Large anchor boxes
M4	RestNet152 + FPN	8 × 10; Single RoI	Customized anchor boxes
M5	SPPF	N/A	N/A

Table 4 The COCO detection evaluation matrix of the models in Table 4

Name	IoU	Size	M1	M2	M3	M4	M5
Average precision	$IoU \geq 0.5$	All	92.3%	71.4%	77.4%	87.5%	90.4%
		Small	78.9%	47.2%	54.9%	64.7%	74.3%
		Medium	94.6%	73.9%	83.5%	90.4%	92.6%
		Large	95.8%	87.0%	86.2%	89.4%	95.5%
	$0.5 \leq IoU \leq 0.95$	All	63.1%	31.2%	36.6%	40.2%	60.4%
		Small	50.6%	15.6%	22.6%	27.5%	48.3%
		Medium	56.7%	29.3%	36.8%	40.2%	55.7%
		Large	67.9%	44.7%	47.4%	47.5%	66.2%
Average recall	$IoU \geq 0.5$	ALL	91.4%	69.1%	77.6%	86.7%	89.2%

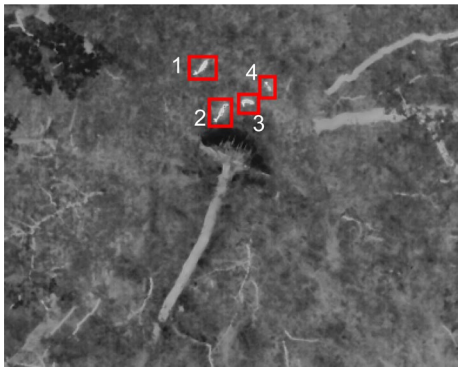


Fig. 15 There are four deer objects in a thermal image. according to the standard in Table 3, Deer 1 and Deer 2 are large objects, Deer 3 is medium, and Deer 4 is small

detections included two false-positive results. Under the condition of same training epochs, models M1, M3 and M4 demonstrate a faster convergence speed compared to M2. These outcomes serve as evidence that the structure of ResNet152 + FPN has more advantages in the extraction of feature information from original images than VGG16. ResNet152 + FPN contributed positively to the convergence efficiency of an object detection model and improve the performance in the detection of small animal objects in low-resolution thermal images.

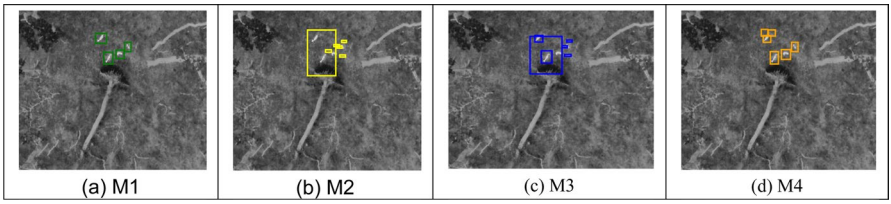


Fig. 16 The detection results from the four models are respectively shown in (**a**), (**b**), (**c**), and (**d**)

M1 and M3 have the same structures, except the strategy of anchors generation. As shown in Fig. 17, the thermal image contains six deer, and the model M1 successfully detected all deer objects, but the model M3 missed two small objects labeled by yellow boxes, which proves that customized anchor boxes are more suitable for small object detection than using hand-picked anchor boxes.

M1 and M4 have the same structures, except the strategy of RoI Pooling. M1 adopts the method of Fusion of Multi-scale RoI Align to construct the feature map of RoIs for species identification, while M4 only uses the output of a RoI Align layer. As shown in Fig. 18, there are 73 deer in the thermal image. 72 deer objects were detected by M1 and M4. In the specific region marked by red arrow as shown in Fig. 18, there are three deer, but only two were detected. However, among the 72 objects, 5 objects were wrongly classified as "Other" by the model M1, and 10 objects were wrongly classified as "Other" by the model M4, which means that the method of Fusion of Multi-scale RoI Align proposed in this paper can improve the Average Precision of species identification of animals in thermal images.

Finally, we compared the model M1 with the model M5, which is constructed based on YOLOv8. As shown in Fig. 19, the thermal image contains 73 deer objects. M1 found 72 deer objects and M5 detected 73 deer objects. More details are shown in Fig. 19c and Fig. 19d, which reveals that M1 missed one, and M5 also missed one, but also detected a false-positive one. From the perspective of species identification, 5 objects were wrongly classified as "Other" by the model M1, and 8 objects

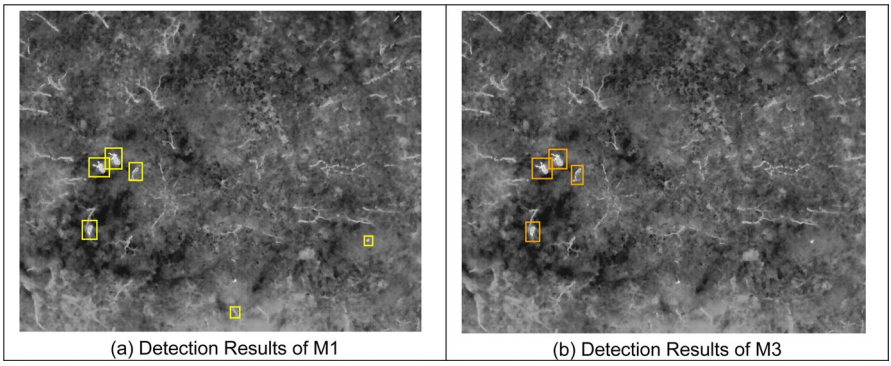


Fig. 17 **a** The detection results from the model M1 are colored by *yellow*. **b** the detection results from the model M3 are colored by *orange*

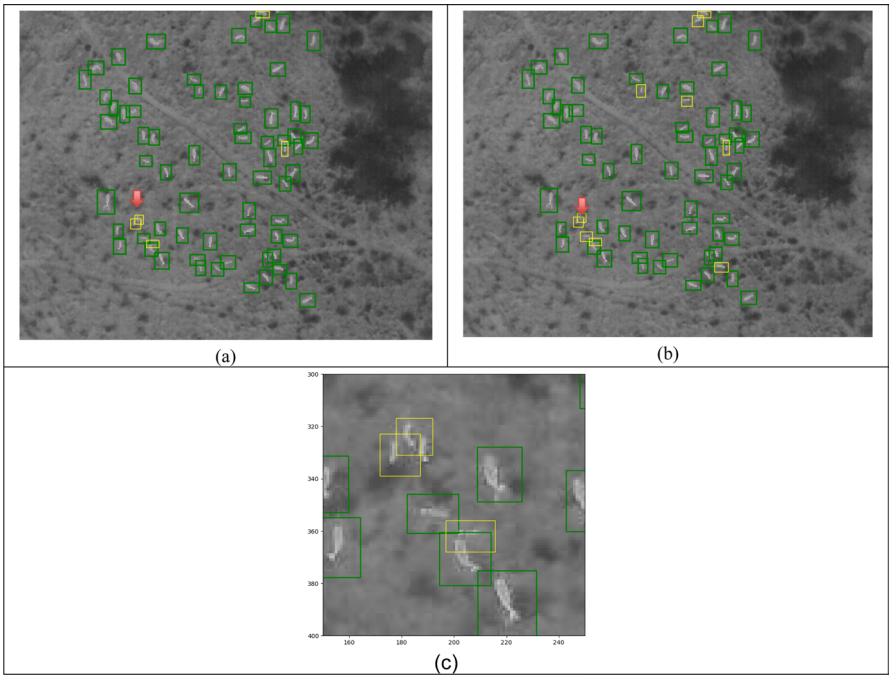


Fig. 18 **a** Species Identification Results of M1. **b** Species Identification Results of M2. **c** The Details of Region Maked by *Red Arrow* in (**a**) and (**b**). The objects labeled by *green boxes* were identified as "Deer", and the objects labeled by *yellow boxes* were identified as "Other"

were wrongly classified as "Other" by the model M5. According to the experimental results, M1 and M5 have similar performance of object detection. However, M1 is slight better in species identification for small objects in low-resolution thermal images.

In addition, wild animals typically exhibit a strong sense of territoriality, so the animals in the UAV thermal images are all of the same species. To verify that model M1 can detect and identify wild deer among other animals, we mosaiced multiple thermal images from the validation dataset into a single image, as shown in Fig. 20a. This image contains 12 deer objects labeled with red boxes and 11 other animal objects labeled with yellow boxes. The predicted results by model M1 are displayed in Fig. 20b. Notably, the model not only detected the 12 annotated deer objects but also identified an additional 9 unlabeled deer objects. These experimental results demonstrate that the model M1 constructed in this study can accurately discriminate deer from other animals in UAV thermal images, exhibiting good generalizability.

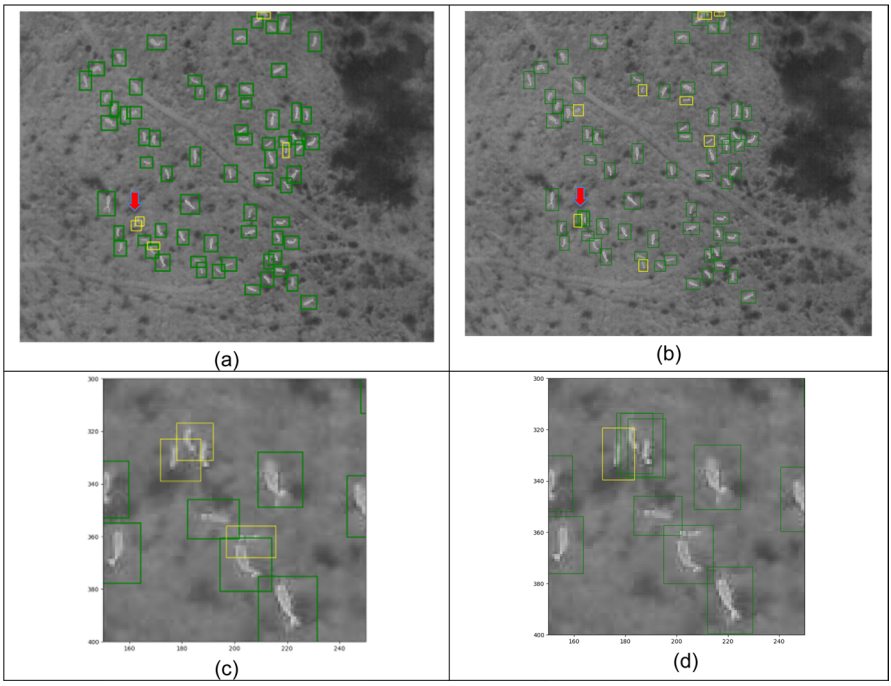
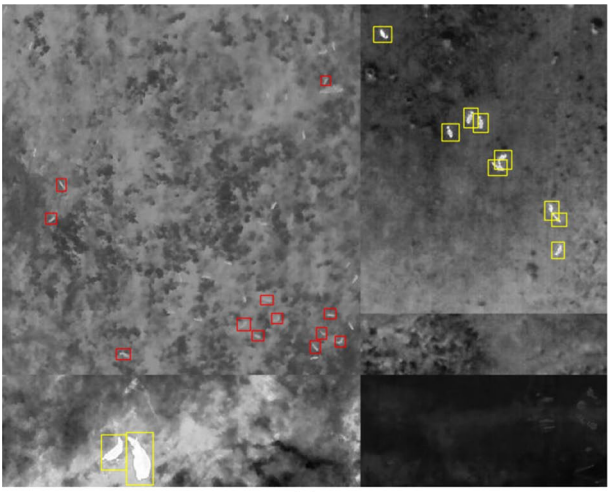


Fig. 19 **a** The detection results of M1; **b** the detection results of M5; **c** the details of the region marked by *red arrow* in (**a**); **d** the details of the region marked by *red arrow* in (**b**)

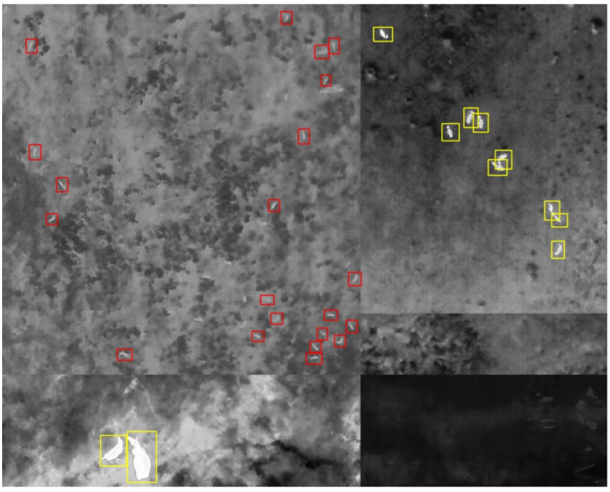
5 Discussion

5.1 Contributions

This research presents a wildlife animal detection model based on Faster R-CNN, specifically designed to effectively detect wild deer from thermal images captured by UAVs, thereby enhancing the efficiency of data collection and analysis during wild deer surveys. The model addresses the challenges associated with detecting small objects in low-resolution UAV thermal images through several tactics. ResNet152 and FPN serve as the backbone, and multiple feature maps are constructed for the Region Proposal Network (RPN). Each feature map enriches the pool of spatial and semantic features available for object localization. The use of customized anchor boxes significantly improves the model's capability to generate robust Regions of Interest (RoIs), tailored for detecting small objects in thermal imagery. Furthermore, the implementation of a fusion of multi-scale RoI Align strategy enhances the extraction of relevant feature information for RoIs, thereby improving the accuracy of wild animal species identification. This research is helpful for effective wild deer monitoring and conservation, providing



(a)



(b)

Fig. 20 **a** The mosaiced image containing deer and other animals; **b** The results predicted by the model M1

valuable insights into deer populations and their behavior in different environmental conditions.

5.2 Future Work

The following three topics are worthy of further exploration.

- Enhancing the precision of species identification for small animal objects in low-resolution thermal images: The target animals in thermal images captured

by UAVs are usually small, which often leads to inadequate feature extraction and subsequent misclassification, even though most are successfully detected. Therefore, designing and constructing advanced deep learning architectures for the feature extraction of small objects contiues to be a critical focus of my future research.

- Enhancing Feature Information in Thermal Images: Thermal images, in comparison to normal RGB images, often have lower resolution, which is not conducive to the detection of small objects. To overcome this limitation, future work should explore methods to enhance the feature information in thermal images. Some professional palettes are designed to improve object features with varying temperatures in thermal images. It is possible to fuse feature information from thermal images with different palettes. Therefore, exploring the utilization of these advantages in future research is recommended.
- When conducting wild deer surveys using drones, the drones capture images at regular intervals while flying back and forth in straight lines across the study areas. To obtain an accurate count of the deer population, it is necessary to set these intervals to ensure a minimum of 15% end overlap and 15% side overlap. Consequently, it is possible for some deer to be captured multiple times. There exists deer object redundancy among the thermal images captured. In the future, we hope to find some methods to find these redundant deer objects in different thermal images to reduce the issue of redundancy.

6 Conclusion

When conducting wild deer surveys in Chitwan National Park, Nepal, the dense coverage of tall trees and vegetation often obscures the presence of wild deer, making it challenging to monitor them with high-resolution RGB cameras. The optimal approach is using UAVs equipped with thermal cameras to monitor wild animals. However, the resolution of thermal images is usually low, and the size of target animals in these images is very small. Current mainstream target detection algorithms cannot be directly applied to detect and identify wild deer in UAV thermal images.

After comparing one-stage YOLOv8 and two-stage Faster R-CNN detection frameworks, we chose to construct an object detection model based on Faster R-CNN for detecting wild deer from UAV thermal images. Specifically, we used a Feature Pyramid Network (FPN) based on ResNet-152 as the backbone to extract feature information from thermal images and construct multi-scale feature maps for object localization. We then set suitable anchor frame sizes and aspect ratios for wild animal detection according to the results of K-means clustering over the collected thermal image dataset. Furthermore, we implemented a fusion of multi-scale RoI Align strategy to enhance the extraction of relevant feature information for RoIs, thereby improving the accuracy of wild animal species identification. The following conclusions were drawn:

1. Compared with the wild animal detection models proposed by some researchers shown in Table 3, the model in this paper achieved better detection performance.

2. The performance of the model proposed in this paper was evaluated using the COCO detection evaluation metric. Under the condition of IoU > 0.5, the results revealed mean Average Precision (mAP) of 92.3% for all objects, 78.9% for small objects, 94.6% for medium objects, and 95.8% for large objects.

Author's Contribution LH, QF, AL, SD, and LR: Conceptualization of research problems. LH, QF: Methodology. QF and LH wrote original draft preparation. LH wrote the main manuscript. LH, AL, SD, and LR. provide investigation and formal analysis. QF, BE and LR thermal image dataset. AL and QF provided funding acquisition. All authors revewed the manuscript.

Funding This research was performed with financial support from the USA National Science Foundation (NSF) grant number BCS-1826839.

Declarations

Conflict of interest The authors declare no conflicts of interest that could influence the outcome or interpretation of this study. The authors declare no competing interests.

References

1. Bengsen, A. J., Forsyth, D. M., Ramsey, D. S. L., Amos, M., Brennan, M., Pople, A. R., Comte, S., & Crittle, T. (2022). Estimating deer density and abundance using spatial mark-resight models with camera trap data. *Journal of Mammalogy, 103*(3), 711–722. https://doi.org/10.1093/jmammal/gyac016
2. Bochkovskiy, A., Wang, C.-Y., & Mark Liao, H.-Y. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv. https://doi.org/10.48550/arXiv.2004.10934.
3. Carl, C., Schönfeld, F., Profft, I., Klamm, A., & Landgraf, D. (2020). Automated detection of European wild mammal species in camera trap images with an existing and pre-trained computer vision model. *European Journal of Wildlife Research, 66*, 1–7.
4. Chabot, D., & Bird, D. M. (2012). Evaluation of an off-the-shelf unmanned aircraft system for surveying flocks of geese. *Waterbirds, 35*(1), 170–174.
5. Chabot, D., Dillon, C., & Francis, C. M. (2018). An approach for using off-the-shelf object-based image analysis software to detect and count birds in large volumes of aerial imagery. *Avian Conservation & Ecology*, *13*(1).
6. Chabot, D., & Francis, C. M. (2016). Computer-automated bird detection and counts in high-resolution aerial images: A review. *Journal of Field Ornithology, 87*(4), 343–359.
7. Choiński, M., Rogowski, M., Tynecki, P., Kuijper, D. P. J., Churski, M., & Bubnicki, J. W. (2021). A first step towards automated species recognition from camera trap images of mammals using AI in a European temperate forest. In: *Computer information systems and industrial management: 20th international conference, CISIM 2021, E\lk, Poland, September 24–26, 2021, Proceedings 20*, 299–310. Springer.
8. Christiansen, P., Steen, K. A., Jørgensen, R. N., & Karstoft, H. (2014). Automated detection and recognition of wildlife using thermal cameras. *Sensors, 14*(8), 13778–13793.
9. Conner, M. M., & McKeever, J. S. (2020). Are composition surveys for mule deer along roads or from helicopters biased? Lessons from the field. *Wildlife Society Bulletin, 44*(1), 142–151.
10. Eikelboom, J. A. J., Wind, J., van de Ven, E., Kenana, L. M., Schroder, B., de Knegt, H. J., van Langevelde, F., & Prins, H. H. T. (2019). Improving the precision and accuracy of animal population estimates with aerial image object detection. *Methods in Ecology and Evolution, 10*(11), 1875–1887. https://doi.org/10.1111/2041-210X.13277
11. Freeman, M. S., Dick, J. T. A., & Reid, N. (2022). Dealing with non-equilibrium bias and survey effort in presence-only invasive species distribution models (iSDM); predicting the range of Muntjac Deer in Britain and Ireland. *Ecological Informatics, 69*, 101683.

12. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In: 770–78. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.
13. Jobin, B., Labrecque, S., Grenier, M., & Falardeau, G. (2008). Object-based classification as an alternative approach to the traditional pixel-based classification to identify potential habitat of the grasshopper sparrow. *Environmental Management, 41*, 20–31.
14. Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., et al. (2022). NanoCode012. "Ultralytics/Yolov5: V7.0 - YOLOv5 SOTA Realtime Instance Segmentation." Zenodo. https://doi.org/10.5281/zenodo.3908559.
15. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.
16. Lautenschlager, R. A. (2021). Deer (Track-Pellet). In: CRC handbook of census methods for terrestrial vertebrates, pp. 249–250. CRC Press.
17. Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. https://doi.org/10.1109/5.726791
18. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2018). Focal loss for dense object detection. arXiv. https://doi.org/10.48550/arXiv.1708.02002.
19. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M., *Computer vision – ECCV 2016*, pp. 21–37. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_2.
20. Lyu, H., Qiu, F., An, Li., Stow, D., Lewison, R., & Bohnett, E. (2024). Deer survey from drone thermal imagery using enhanced faster R-CNN based on ResNets and FPN. *Ecological Informatics, 79*(March), 102383. https://doi.org/10.1016/j.ecoinf.2023.102383
21. Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences, 115*(25), E5716–E5725.
22. Padilla, R., Netto, S. L., & da Silva, E. A. B. (2020). A survey on performance metrics for object-detection algorithms. In: *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 237–242. https://doi.org/10.1109/IWSSIP48289.2020.9145130.
23. Peng, J., Wang, D., Liao, X., Shao, Q., Sun, Z., Yue, H., & Ye, H. (2020). Wild animal survey using UAS imagery and deep learning: modified faster R-CNN for Kiang Detection in Tibetan Plateau. *ISPRS Journal of Photogrammetry and Remote Sensing, 169*(November), 364–376. https://doi.org/10.1016/j.isprsjprs.2020.08.026
24. Podobna, Y., Sofianos, J., Schoonmaker, J., Medeiros, D., Boucher, C., Oakley, D., & Saggese, S. (2010). Airborne multispectral detecting system for marine mammals survey. In: *Ocean Sensing and Monitoring II*, 7678:136–44. SPIE.
25. Rangdal, M. B., & Hanchate, D. B. (2014). Animal detection using histogram oriented gradient. *International Journal on Recent and Innovation Trends in Computing and Communication, 2*(2), 178–183.
26. Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: towards real-time object detection with region proposal networks. arXiv. https://doi.org/10.48550/arXiv.1506.01497
27. Rush, G. P., Clarke, L. E., Stone, M., & Wood, M. J. (2018). Can drones count gulls? Minimal disturbance and semiautomated image processing with an unmanned aerial vehicle for colony-nesting seabirds. *Ecology and Evolution, 8*(24), 12322–12334.
28. Schoonmaker, J. S., Podobna, Y., Boucher, C. D., Statter, D. R., & Contarino, V. M. (2011). Electro-optical approach for airborne marine mammal surveys and density estimations. *US Navy Journal of Underwater Acoustics, 61*(4), 968–985.
29. Selby, W., Corke, P., & Rus, D. (2011). Autonomous aerial navigation and tracking of marine animals. In: *Proceedings of the Australian Conference on Robotics and Automation (ACRA)*.
30. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv. https://doi.org/10.48550/arXiv.1409.1556.
31. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going deeper with convolutions. arXiv. https://doi.org/10.48550/arXiv.1409.4842.
32. Torney, C. J., Dobson, A. P., Borner, F., Lloyd-Jones, D. J., Moyer, D., Maliti, H. T., Mwita, M., Fredrick, H., Borner, M., Grant, J., & Hopcraft, C. (2016). Assessing rotation-invariant feature classification for automated wildebeest population counts. *PLoS ONE, 11*(5), e0156342.

33. Vecvanags, A., Aktas, K., Pavlovs, I., Avots, E., Filipovs, J., Brauns, A., Done, G., Jakovels, D., & Anbarjafari, G. (2022). Ungulate detection and species classification from camera trap images using retinanet and faster R-CNN. *Entropy, 24*(3), 353.
34. Wang, C.-Y., Bochkovskiy, A., & Mark Liao, H.-Y. (2022). YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv. https://doi.org/10.48550/arXiv.2207.02696.
35. Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. arXiv. https://doi.org/10.48550/arXiv.1611.05431.
36. Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023).“Object detection in 20 years: a survey. In: *Proceedings of the IEEE.*

Authors and Affiliations

Haitao Lyu[1] · Fang Qiu[1] · Li An[2] · Douglas Stow[3] · Rebecca Lewision[3] · Eve Bohnett[4]

✉ Fang Qiu
ffqiu@utdallas.edu

[1] Geospatial Information Science, The University of Texas at Dallas, 800 West Campbell Road, Richardson, TX 75080, USA

[2] International Center for Climate and Global Change Research, Complex Human-Environment Systems Laboratory, College of Forestry, Wildlife, and Environment, Auburn University, Auburn, AL, USA

[3] Department of Geography, San Diego State University, San Diego, CA 92182, USA

[4] Department of Landscape Architecture, College of Design Construction and Planning, University of Florida, Gainesville, FL, USA