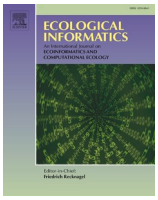
Ecological Informatics 79 (2024) 102383

Contents lists available at ScienceDirect

Ecological Informatics

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



Deer survey from drone thermal imagery using enhanced faster R-CNN based on ResNets and FPN

Haitao Lyu [a], Fang Qiu [a,*], Li An [b], Douglas Stow [c], Rebecca Lewison [d], Eve Bohnett [d,e]

[a] Geospatial Information Science, The University of Texas at Dallas, Richardson, TX, USA
[b] International Center for Climate and Global Change Research, Complex Human-Environment Systems Laboratory, College of Forestry, Wildlife, and Environment, Auburn University, Auburn, AL, USA
[c] Department of Geography, San Diego State University, San Diego, CA, USA
[d] Department of Biology, San Diego State University, San Diego, CA, USA
[e] Department of Landscape Architecture, College of Design Construction and Planning, University of Florida, Gainesville, FL, USA

ARTICLE INFO

Keywords:
UAV
Faster R-CNN
ResNet
FPN
Thermal image
Small object detection

ABSTRACT

Deer surveys play an important role in the estimation of local ecological balance. In the Chitwan National Park of Nepal, the dense tree canopies and tall vegetation often obscure the presence of wild deer, which has a negative effect on the accurate population surveys of wild deer. UAVs equipped with infrared sensors have been increasingly used to monitor wild deer by capturing a lot of images. How to automatically recognize and obtain the number of deer objects from thermal images is becoming an important research topic. Due to the difference between thermal images and true-color images, as well as the variations in deer object sizes in these two types of images, current ready-to-use object detection models, designed for true-color imagery, are ill-suited for the task of detecting small deer objects within thermal imagery. In this paper, an enhanced Faster R-CNN was constructed to detect small deer objects from thermal images, in which a Feature Pyramid Network (FPN) based on a residual network is used to improve feature extraction for small deer objects and multi-scale feature map constrution for the subsequent region proposals searching, bounding box regression, and regions of interest (RoIs) classification. In addition, small-scaled anchor boxes and a multi-scale feature map selection criterion are devised to improve the detection accuracy of small objects. Finally, based on Faster R-CNN, FPN, and different residual networks including ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152, we constructed five object detection models, and evaluated their detection performance by using COCO evaluation matrix. Under the condition of $IoU \geq 0.5$, the integration of Faster R-CNN, FPN, and ResNet18 demonstrated to perform better than others. Specifically, The COCO evaluation results revealed an Average Precision (AP) score of 91.6% for all deer objects. Small deer objects (area $\leq$ 200 pixels) achieved an AP score of 73.6%, medium deer objects (200 $<$ area $\leq$ 400 pixels) demonstrated an AP score of 93.4%, and large deer objects (area $>$ 400 pixels) achieved the highest AP score of 94.3%. Our research is helpful for effective wild deer monitoring and conservation and can be a valuable reference for the exploration of small object detection from low-resolution thermal images.

1. Introduction

Deer survey refers to the process of determining the number of deer and their distribution in a specific area or habitat (Bengsen et al., 2022a; Schwarz and Seber, 1999). Deer surveys are often conducted within a particular ecosystem, such as a forest, a wildlife reserve, or a national park. By accurately estimating the number of deer in a given area, wildlife authorities and conservationists not only can have a better understanding of the distribution of deer population, but also provide an important barometer for the local ecological balance of an ecosystem based on the interactions of deer with other species (Forsyth et al., 2022). When deer population decreases a lot in a specific area, it often signals an increase in carnivorous animals or rampant poaching. Conversely, a surge in deer population may indicate a decline in carnivorous species or an overabundance of deer. Therefore, accurate, detailed, and up-to-date wild deer surveys are of great benefit to wildlife

* Corresponding author.
E-mail addresses: hxl170008@utdallas.edu (H. Lyu), ffqiu@utdallas.edu (F. Qiu), anli@auburn.edu (L. An), stow@sdsu.edu (D. Stow), rlewison@sdsu.edu (R. Lewison), evebohnett@ufl.edu (E. Bohnett).

https://doi.org/10.1016/j.ecoinf.2023.102383
Received 28 July 2023; Received in revised form 17 October 2023; Accepted 16 November 2023
Available online 25 November 2023

management and conservation.

Traditional wild deer surveys were often conducted using ground-based counting by field investigators who observe wild animals at close range. For example, the population of wild deer is estimated by tracking and counting deer and their pellets (Lautenschlager, 2021). Ground-based counting is an arduous and exhausting work for field inspectors, and the survey may not be accurate due to deer's elusive behavior, dense vegetation, varied landscape, and presence of dangerous carnivorous species. Wild deer is known for being very alert and tend to avoid human presence, making it very difficult for surveyors to get close enough for clear observations (Freeman et al., 2022). Tall grass and thick tree branches and canopies often obscure visibility and make it hard to spot and count deer. Deer may be distributed in grasslands, wetlands, and forests, each with their unique challenges to access. Additionally, various dangerous carnivorous species in these landscapes are potential risks to surveyors.

Camera trapping is also a popular method used in wild deer surveys. Camera traps are strategically placed in various locations, such as trails, watering holes, and other areas where deer are likely to pass. Fig. 1 (a) shows a deer captured by a camera trapping when it passed by the trap. Camera trapping can generate a lot of images and automatic object detection models can be used to alleviate the task of reviewing images. Four deer species were surveyed with camera trappings in the state of Queensland, New South Wales, and Victoria, eastern Australia, and a detection model was used to count the number of deer in the camera trapping images, and to estimate the density of wild deer (Bengsen et al., 2022b). Compared to ground-based counting, camera trapping allows researchers to study wildlife without direct human presence, reducing disturbance to the animals, and can operate for extended periods. However, camera traps are stationary devices, and their deployed locations are often chosen based on human subjective judgement or prior knowledge. This can introduce bias, resulting in certain areas or species to be overrepresented, while others underrepresented or missed entirely.

In recent years, remote sensing technologies have been embraced in wild deer surveys, offering a powerful and efficient means to observe and monitor wild deer populations and their habitats over large geographic areas. Remote sensing-based surveys may be conducted with manned aircraft census and satellite monitoring. For example, a manned helicopter was used to monitor and estimate the population of wild deer in the Sierra Nevada as shown in Fig. 1 (b) (Conner and McKeever, 2020). While manned aircraft surveys offer flexibility regarding survey timing and area, they also come with a relatively high cost to pay for not only the aircraft but also the qualified and skilled pilots, and may exert disturbance to the animals due to noise imposed by the flying aircraft (Petso et al., 2021). Satellite-based wildlife monitoring, capable of counting wild animals from space, primarily relies on very-high-resolution ($\leq$1 m) satellite imagery. For instance, the population of Weddell seals in the coast of Antarctica was estimated using high-resolution satellite images . These methods offer extensive observation coverage, short revisit intervals, and minimal disturbance to the animals. Nonetheless, it's necessary to note that even with very-high-resolution satellite imagery, these techniques are only effective in recognizing larger individual animals (such as wildebeests shown in Fig. 1 (c)), but not smaller animals such as wild deer.

With the recent advancement of drone technology and the decreasing of the equipment costs, unmanned aerial vehicles (UAVs) have emerged as a promising alternative for conducting wildlife surveys. Unlike manned aircraft requiring highly skilled pilots, UAV surveys can be operated by average researchers with moderate training and therefore is much cost-effective (Nazir and Kaleem, 2021). Furthermore, UAVs offer greater flexibility in revisiting survey areas at any specific time, compared to the fixed date and time of satellite remote sensing data acquisition. Additionally, UAVs can be configured or customized with different types of sensors, allowing them to capture not only very high-resolution true-color imagery but also thermal imagery, enhancing the versatility to perform wildlife surveys at the landscapes covered by dense forest and tall vegetation. Fig. 2 shows two images simultaneously captured by a UAV in an area with 8 wild deer using two different sensors. Fig. 2 (a) is a true-color image with a resolution of 8000 $\times$ 6000 pixels, with 4 deer in the open space near the road (labeled by green rectangles) easily seen, two deer between the tree canopies (labeled by yellow rectangles) visible and not easily be detected, and two deer covered partially or completely by tree canopies (labeled by red rectangles) not visible. Fig. 2 (b) is a thermal image with a resolution of 640 $\times$ 512 pixels, with 8 wild deer all visible, even the two deer beneath the tree canopies, attributing to their body temperature being higher than the ambient background. Therefore, UAVs are increasingly employed for wildlife monitoring and counting, and researchers have begun to adopt thermal sensors for wild deer surveys (Preston et al., 2021). The focus of study is to monitor wild deer in the Chitwan National Park of Nepal. Most parts in the park are covered by trees and tall grasses, which make it difficult for true-color sensors to detect wild deer due to the occlusion of vegetation canopies as shown in Fig. 2 (a). Therefore, thermal cameras onboard the Mavic 2 Enterprise Advanced DJI Drones are used as the primary sensors for wild deer survey in this research.

UAV surveys and camera trapping can easily capture thousands of images, resulting in huge image datasets that require careful examination for the identification of wild animals. Manual recognition and counting of animals from the imagery is relatively accurate and reliable. For example, white-tailed deer in two United States National Parks (Harpers Ferry National Historic Park and Monocacy National Battlefield) surveyed by UAVs equipped with thermal sensors were enumerated manually in order to estimate their density (Preston et al., 2021). However, manual recognition and counting requires great effort to review a massive number of images, which is a labor-intensive and time-consuming task (Greenberg et al., 2019). To overcome this difficulty, automatic or semi-automatic methods based on artificial intelligence have been used to detect animals from images. Deep learning models characterized by convolutional neural networks (CNNs) are being increasingly used to automate the animal recognition and counting tasks, and their performance is continuously improved with the introduction of enhanced architecture (Kaur and Singh, 2022). For example,

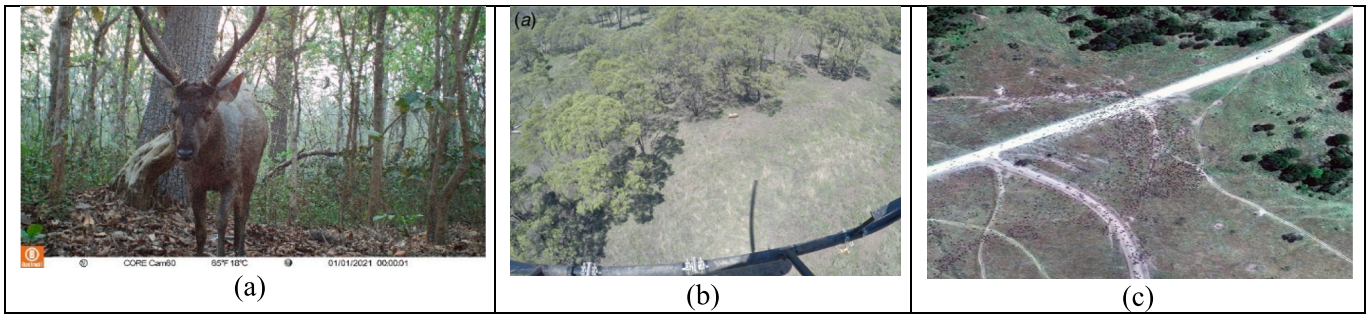


Fig. 1. (a) a deer was captured by a camera trapping; (b) a deer was monitored by a helicopter; (c) wildebeest migration captured by GeoEye-1 Satellite (Copyright: MAXXAR www.satimagingcorp.com/).

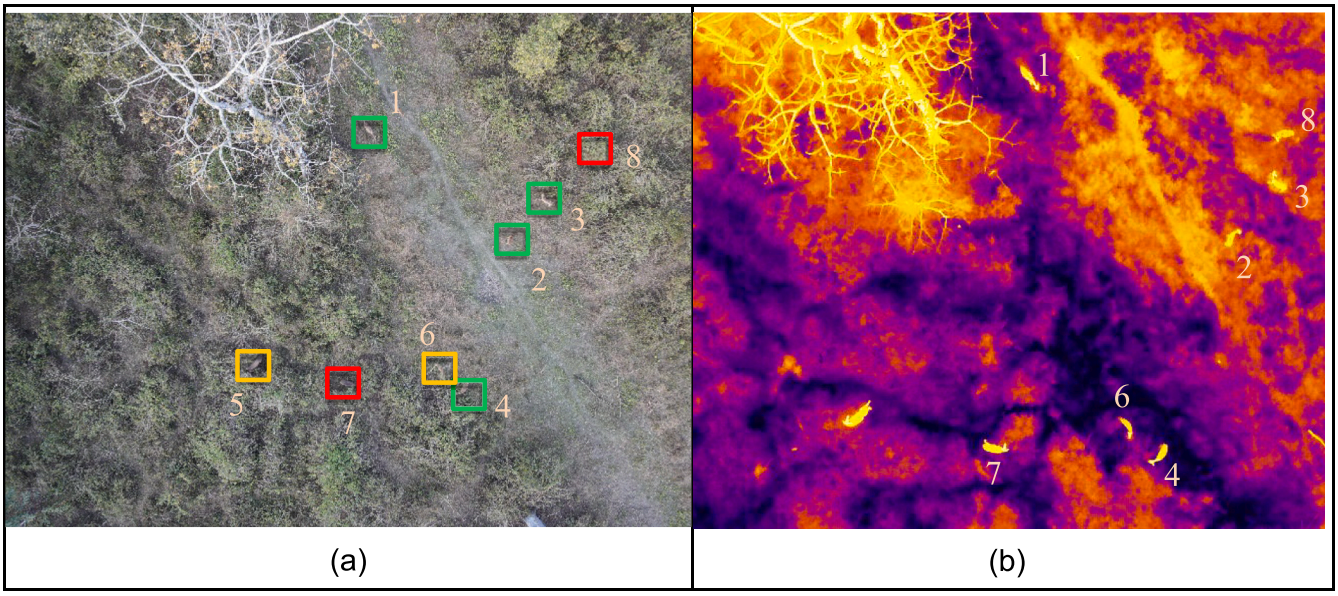


Fig. 2. Two wild deer were not seen on the 8 K true-color image (left) due to canopy cover, but they were visible on the thermal image (right) due to their higher body temperature.

a deep neural network architecture based on ResNet 50 was employed to detect wild animals from the Snapshot Serengeti true-color imagery captured by camera trapping in Africa, and achieved a detection accuracy of 93.8% (Norouzzadeh et al., 2018). A model based on Feature Pyramid Network (FPN) and ResNet50 was applied to detect elephants, giraffes, and zebras from high-resolution true-color UAV images captured over an open savanna in the Tsavo National Park of Kenya, and achieved a detection accuracy of 95% for elephant, 91% for giraffes, and 90% for zebras (Eikelboom et al., 2019). However, their method depends on using a fix-size sliding window to scan images for object detection. As the window slides across an image, the same region of the image may be evaluated multiple times with slight positional variations, leading to prolonged processing time, especially for large images. In addition, when dealing with objects of varying sizes using a fix-size sliding window, the detection precision for objects smaller than the window size tends to decline (Eikelboom et al., 2019). To address these problems, object detection architecture with varied window sizes such as Faster R-CNN and YOLO have been adopted to detect objects. For example, a pre-trained Faster RCNN + InceptionResNetV2 model was utilized to detect European mammals from camera trapping imagery and achieved a detection accuracy of 94% (Carl et al., 2020). A RetinaNet and a Faster R-CNN + ResNet50 were utilized to detect ungulate animals including deer and boars from camera trapping imagery (Vecvanags et al., 2022). However, there is a relative scarcity of research employing these improved models for detecting animals from UAV images, with an exception by (Peng et al., 2020) who employed Faster R-CNN to detect kiang objects from true-color drone images with a resolution of 6000 × 4000 pixels, achieved an overall precision of approximately 90%.

To the best of our knowledge, detecting small wild animals like deer from UAV thermal images using these improved object detection models has not been found in the literature. One possible reason is likely that the size of deer objects in pixels in a UAV thermal image is much smaller than that in a high-resolution true-color image. The size of deer objects in pixels in a true-color image (Fig. 2 (a)) and those in a thermal image (Fig. 2 (b)) are listed in Table 1. On average, the total image size of deer objects in the thermal image is about 100 times smaller than that in the true-color image, providing very limited feature information to represent deer objects. Despite their outstanding performance for animal detection in true-color imagery, current ready-to-use models designed for true-color camera trapping or UAV imagery could not be directly applied to UAV thermal images for deer detection. Therefore, customizing the improved deep learning structure and finding optimal model configuration suitable for detecting small objects such as deer from UAV thermal images is worthy of further investigation, which comprises the main purpose of this study.

For deep learning-based object detection, two fundamental elements, feature maps and anchor boxes, play critical roles. Feature maps are extracted from original images and utilized to identify potential objects of interest and locate their rough positions within an image. Subsequently, anchor boxes come into play for generating regions of interest (RoIs) and facilitating bounding box regression, which enables the refinement of RoIs to better match the bounding box of the objects of interest. In the general definition of CNNs, the CNN layers that are closer to the input layer are called shallower layers, while deeper layers are those more distant from the input layer. In the architecture of object detection, various CNNs—such as ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152—differ in the overall depth of their layers. These CNNs are used to extract feature maps from the original imagery. Deeper layers derive additional features from the output of shallower layers through progressive downscaling. Consequently, feature maps from shallower layers contain more spatial feature information due to higher resolution, while those from deeper layers have more bands and may provide more abstract semantic feature information. However, when dealing with small objects containing only a limited number of pixels, the vital spatial information about these objects can potentially be lost in the deeper layers as part of the downsizing process, which may not only fail to contribute to detection process but can also diminish detection precision. Additionally, to ensure that the predicted RoIs align

Table 1
the size of deer objects (Width, Height) in Fig. 2.

	Deer 1	Deer 2	Deer 3	Deer 4	Deer 5	Deer 6	Deer 7	Deer 8
True Color	(230,314)	(211,216)	(240,245)	(230,314)	(309,343)	(220,314)	(323,300)	(265,260)
Thermal	(23,30)	(20,21)	(23,24)	(23,28)	(30,34)	(19,31)	(31,30)	(26,22)

well with the bounding boxes of objects of various sizes, it is essential that anchor box sizes are not fixed but instead set to be comparable to those of the objects of interest. For small object detection, an anchor box that is too large may introduce unwanted background information, while one that is too small may only encompass part of the object, both of which can negatively impact the process of object identification and bounding box regression.

The purpose of this research is to address the above-mentioned problems of small object detection based on deep learning and to achieve optimal model structures and configurations for deer surveys using UAV thermal images. For this purpose, an enhanced Faster R-CNN based on FPN + ResNets was constructed, and the following three objectives will be achieved in this paper. Firstly, Feature Pyramid Network (FPN) and residual networks are used to construct multi-scale feature maps. Specifically, residual networks are utilized to extract feature information from original imagery. The features obtained from both shallower and deeper layers are subsequently fused by FPN to generate diverse feature maps with different resolutions. Secondly, customized anchor boxes that match with deer of different sizes in the UAV thermal images are adopted to improve the precision of small object detection. Thirdly, the multiscale feature map selection criterion is defined, allowing the model to generate RoIs. These RoIs draw upon feature information from the respective multiscale feature maps based on their sizes, which can contribute to the efficiency of object identification and facilitate precise bounding box regression. Finally, the model proposed in this paper was tested in UAV thermal imagery including 2278 thermal images and 13,509 deer instance annotations. Based on the COCO evaluation matrix, the model obtained an Average Precision (AP) score of 91.6% for all deer objects. Specifically, small deer objects (area $\leq$ 200 pixels) achieved an AP score of 73.6%, medium deer objects (200 $<$ area $\leq$ 400 pixels) demonstrated an AP score of 93.4%, and large deer objects (area $>$ 400 pixels) achieved the highest AP score of 94.3%.

The rest of the paper is organized as follows. In Section 2, an overview of using deep learning to detect objects in wild animal surveys is introduced. This section offers insights into the existing research in this domain. Section 3 delves into the detailed description of our small deer object detection model based on Faster R-CNN, FPN, and ResNets. This section outlines the modifications made to the original model and highlights the novel techniques employed to improve the detection of small deer objects in low-resolution thermal images. In Section 4, the experimental results and analysis are presented. These experiments can provide evidence to support the efficacy of our approaches. Finally, Section 6 concludes the paper, summarizing the key findings and contributions. This section also discusses some future work in this field.

2. Related work

Wildlife surveys, crucial for understanding biodiversity and population trends, often involve the collection of a vast number of images. This abundance of visual data presents a significant challenge as the manual review and identification of animals within these images are labor-intensive and time-consuming. To address this issue, various automatic and semiautomatic methods were proposed to detect animals from images. From the perspective of techniques, these methods can be categorized into two classes: pixel-based classification using machine learning and region-based classification using deep learning.

Pixel-based classification methods, such as supervised classification, unsupervised classification, and threshold setting, are the most common methods for detecting animals in remote sensing images (Peng et al., 2020). For example, threshold setting is a simple and widely used approach in wildlife detection 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 all regions or pixels that surpass this threshold as potential animal regions. In (Jobin et al., 2008), 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. While these approaches work well for targets with distinct gray values that significantly differ from the background, they often exhibit limited accuracy in complex environments where animals blend with their surroundings. Subsequently, more stable hand-crafted features, such as Histogram of 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 (Rangdal and Hanchate, 2014). (Torney et al., 2016) introduced a method that combined rotation-invariant object descriptors with machine learning algorithms to detect wildebeests from aerial images. This approach yielded better results compared to manual operations. (Rey et al., 2017) proposed a semi-automatic system for detecting large mammals in UAV imagery with a high recall rate. But these methods often lacked robustness and still struggled with complex backgrounds.

Recently, with the breakthrough in deep learning, particularly the development of CNNs has revolutionized the field of animal detection. With CNNs, feature extraction became more automated, enabling the models to learn hierarchical representations from raw image data. One of the most popular methods was to use image classification CNNs to find potential areas containing animals by using the sliding window approach across the image. In (Barbedo et al., 2019; Barbedo et al., 2020), famous models, such as VGG, ResNet, AlexNet, GoogleNet, DenseNet, and NASNet, were used to detect cattle from 4 K-resolution images. (Kellenberger et al., 2018) constructed a CNN model based on Resnet-18 to detect large mammals from a dataset acquired over the Kuzikus wildlife reserve in eastern Namibia and got a high recall up to 90%. However, this method is very time-consuming and computationally expensive. Later, the methods using the sliding window techniques were replaced by object detection models based on deep learning, such as R-CNN (Bharati and Pramanik, 2020), Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2016), YOLO (Bochkovskiy et al., 2020), and RetinaNet (Lin et al., 2018). These models combined region proposal techniques with deep CNNs, significantly improving detection accuracy and efficiency. Researchers started applying these models to animal detection tasks. (Eikelboom et al., 2019) utilized RetinaNet to detect elephants, giraffes, and zebras from aerial images in Kenya, and they obtained the accurate ratio of 95% for elephants, 91% for giraffes and 90% for zebras. (Aburasain et al., 2021) used a single-pass deep CNN known as YOLOv3 to detect cattle from drone images and got a F-score of 0.93. In (Popek et al., 2023), a Faster R-CNN model was used to detect deer from the images from camera traps, and got an accuracy of 0.87. Even though these models significantly improve the effectiveness and efficiency of animal detection, the detection accuracy is not stable, especially when dealing with imagery containing objects of varying sizes or objects vary significantly in size within an image. Feature Pyramid Network (FPN) proposed by (Lin et al., 2016) was used to address this problem. The key idea is to leverage the inherent multi-scale, pyramidal hierarchy of deep convolutional neural networks (DCNNs) to construct feature pyramids, which are used to detect objects at different scales. Based on different DCNNs, various FPNs have been constructed. In the research area of animal detection, FPN based on ResNet50 (ResNet50FPN) is widely used as the backbone of object detection models for wildlife surveys (Nazir and Kaleem, 2021). For example, ResNet50FPN was used to remotely detect sick chicken from a poultry farm and obtained a detection accuracy of 93.7% (Zhang and Chen, 2020). A Faster R-CNN integrating ResNet50FPN was constructed to detect big animals from Google Open Images and COCO datasets, such as Bear, Fox, Dog, Horse, Goat, Sheep, Cow, Zebra, Elephant, and Giraffe. They got a mean average precision of 0.81(Yudin et al., 2019). In (Delplanque et al., 2022), FPN based on ResNet101 was used to generate feature maps for object detection models to to detect six types of African mammals of Topi, Buffalo, Kob, Warthog, Waterbuck, and Elephant, and got a mean average precision of 0.82.

Currently, most above research mainly focuses on detecting larger objects from high-resolution true-color images, and some models have

obtained the outstanding performance of animal detection. However, the research of detecting small objects in thermal imagery remains relatively unexplored. Small objects in thermal images captured by drones may appear at varying distances from the drone camera, often leading to size variations in the captured images. For example, the deer objects in the UAV thermal imagery used in this paper exhibit a size variation ranging from 15 × 15 pixels to 65 × 65 pixels. It is relatively difficult for current ready-to-use models trained on true-color imagery to be directly applied to detect small objects from UAV thermal images due to the distinct context difference between thermal images and true-color images, as well as limited pixels for representing small objects in UAV thermal images. In addition, ResNet50 has been used to extract feature maps for object detection models by default in many papers. However, as an image progresses through DCNNs, down sampling operations such as pooling reduce the spatial dimensions of the feature maps. This reduction in spatial information is beneficial for capturing larger objects but can be detrimental to the representation of small objects, which leads to the disappearance of fine-grained details, particularly in small objects. Therefore, in this paper, Apart from ResNet50, ResNet18, ResNet34, ResNet101, and ResNet152 are all tested to extract features from UAV thermal images in order to find optimal model structures and configurations for deer survey using UAV thermal images.

3. Methodology

3.1. Overview

Generally, an object detection model takes images as input, and a DCNN, served as the backbone network, is used to construct feature maps. Subsequently, a region proposal network comes into play, and generates region proposals, assigning a probability for containing an object to each region. The derived region proposals are reshaped by a pooling layer to generate Regions of Interest (RoIs). Finally, classification and bounding box regression is engaged to predict both the presence and location of objects within the original images. These types of models are commonly called 'two-stage detectors' due to their two-step process (Goyal et al., 2023).

Faster R-CNN is a two-stage detector widely recognized for its effectiveness in object detection from images. It comprises two main components: (1) a Region Proposal Network (RPN), responsible for generating a set of region proposals; and (2) a Fast R-CNN (Girshick, 2015) module, which classifies all regions into objects or background and refines the boundaries of the objects. Notably, these two model components share common parameters in the convolution layers used for feature extraction, enabling the two components to be trained at the same time to achieve competitive object detection performance. Fig. 3 shows our Faster R-CNN model designed to detect small deer objects from thermal images. In the Faster R-CNN model, feature maps extracted from images play a crucial role because they provide essential spatial and semantic feature information for the RPN to predict RoIs, and some of these predicted RoIs may correspond to background regions. Specifically, according to their positions within original images, the corresponding feature information of the RoIs is obtained from feature maps. Then the classifier based on Fast R-CNN utilizes this feature information to classify the RoIs into deer objects and backgrounds.

The significance of feature maps in object detection is well-acknowledged, but equally important are anchor boxes. Anchor boxes serve as reference templates at different scales and aspect ratios, guiding the object detection process to precisely locate and classify objects of different sizes and shapes. By aligning the predicted bounding boxes with the anchor boxes, the model can detect objects effectively, especially the small ones that may otherwise be overlooked. Inaccurate anchor boxes can impede the model's capability to detect small objects, potentially causing them to be entirely missed during the detection process. Properly selected anchor boxes are indispensable for the subsequent bounding box regression, which refines initial predicted bounding boxes to better fit the objects' actual locations, thus significantly enhancing the model's precision. Anchor boxes also play a role in hard negative mining, which focuses on challenging negative examples during the training process. By selectively mining hard negative examples, the model can learn to better distinguish between objects and background regions, leading to improved overall performance.

3.2. Multi-scale feature map construction

Feature maps are generated by applying a convolutional layer to the input image or the feature map output of the prior layers. For an object detection model, feature maps are utilized to locate the positions of objects and classify them into specific classes. The original Faster R-CNN in (Ren et al., 2016) adopted VGG16 (Simonyan and Zisserman, 2015) as its backbone network to extract feature information from input images. Specifically, the output of a convolutional layer within the VGG16 was

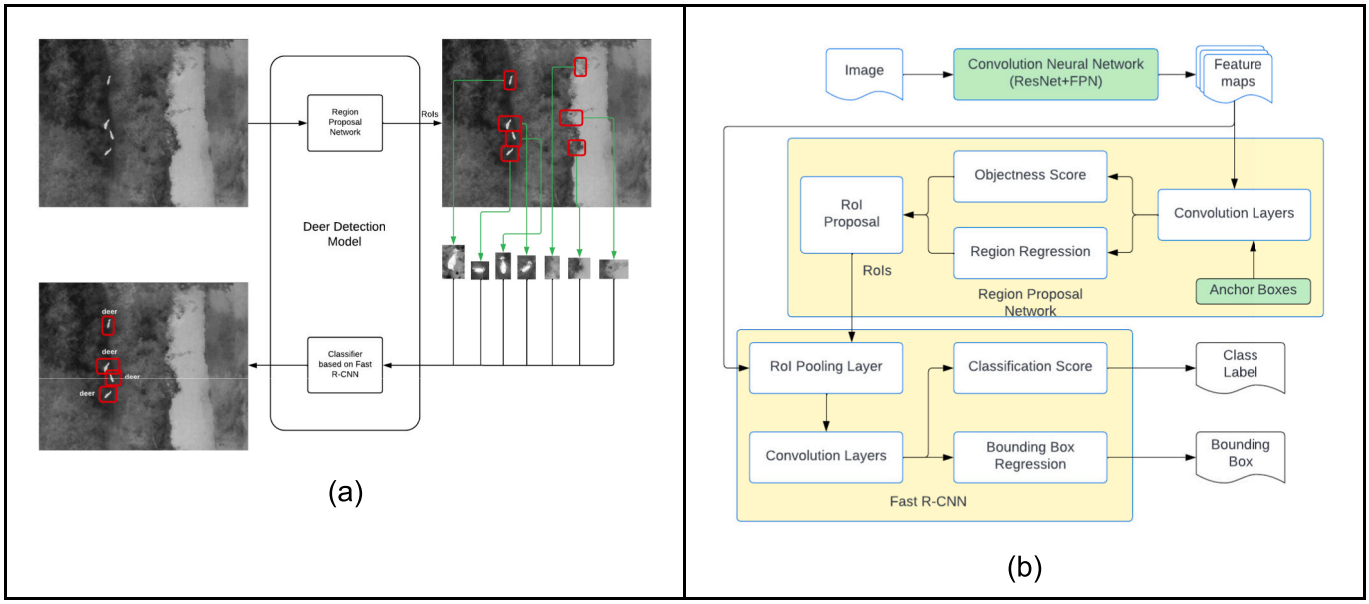


Fig. 3. The structure and flowchart of an enhanced Faster R-CNN in this paper.

utilized as a feature map for further analysis and processing. According to the description of feature maps in (Ren et al., 2016), a thermal image with a resolution of 640 × 512 would result in a feature map with a resolution of 40 × 32. The average small object size in thermal images in this paper is around 30 × 30 pixels. Through the process of feature map extraction, its feature information is condensed to 2 × 2 pixels, resulting in a significant deficiency of spatial feature information necessary for accurate detection of small objects in thermal images. To address this problem, one commonly employed approach is to select a convolutional layer in a CNN model that can produce high-resolution feature maps. However, high spatial feature information in a feature map often pocesses low semantic feature information. A feature map with low semantic features has negative effects on bounding box regression and object classification. To overcome this issue, Feature Pyramid Network (FPN) was used to fuse feature information extracted from different CNN layers. Compared with VGG16, Residual networks exhibit superior capabilities in feature extraction and hierarchical feature representation. Residual networks can effectively generate feature maps at various scales, contributing to a more nuanced and comprehensive understanding of the feature information of small objects. Therefore, the integration of FPN and residual neural networks is used in this paper. Specifically, residual neural networks are used to generate feature maps in different scales. 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 family of residual neural networks includes ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152 according to their number of CNN layers. "Stage" is an important term in the context of residual neural networks, which refers to a specific set of convolution layers that output feature maps with different resolutions. Take ResNet152 as an example in (Fig. 4), it consists of five stages (Conv1, Layer1, Layer2, Layer3, and Layer4) and each stage can produce a feature map. The spatial resolution of each feature map is progressively reduced by a factor of 2, while the number of bands is simutanously increased by a factor of 2. Usually, the traditional FPN neglects the feature maps from Conv1, Layer1 and Layer 2 and only incorporates the two feature maps from Layer 3 and Layer 4. For the detection of large objects from high-resolution true-color images, this may work well. However, for the small objects in UAV thermal images, the spatial feature information left in the two feature maps produced by Layer 3 and Layer 4 may be deficient for accurate detection of small objects. Therefore, all five feature maps produced by Conv1, Layer1, Layer2, Layer3, and Layer4 are utilized in the creation of the final multi-scale feature maps through FPN in this research. Notably, the feature maps extracted from Conv1 and Layer1 exhibit higher spatial resolution and retain more valuable information related to small objects compared to the other feature maps. The inclusion of the feature maps from Conv1, Layer1, and Layer2 allows the model to obtain additional spatial information from their outputs, enhancing the accuracy of small object detection in thermal images.

For example, the structure of FPN based on ResNet152 (ResNet152FPN) is shown in Fig. 5. The bottom-up pathway involes generateing output feature maps from various stages of the network (Lin et al., 2016), which are denoted as $\{C_1, C_2, C_3, C_4, C_5\}$. To merge features extracted from different stages along top-down pathway, the $C_i (i = 1, 2, 3, 4)$ map 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-resolution details from the upsampled map with the existing features. In order to mitigate the potential aliasing artifacts resulting from the merging operation, a 3×3 convolutional operation is applied to each merged map to generate the final feature map. For example, C_5 is the output of Layer4 and is upsampled by a factor of 2 denoted by D_5. The output of Layer3 undergoes a 1×1 convolutional layer to reduce its channel dimensions to be same with C_5 denoted by E_4. By element-wise addition, M_4 is generated and satisfies with the equation of $M_4 = D_5 + E_4$. Then, by a 3×3 convolutional operation, the feature map P_4 is created. Then, M_4 is downsampled by a factor of 2 to be D_4. The output of Layer2 undergoes a 1×1 convolutional layer to generate E_3. By element-wise addition, M_3 is generated and satisfies with the equation of $M_3 = D_4 + E_3$. Similarly, M_2 and M_1 can be generated and satisfy the following equations: $M_2 = D_3 + E_2$ and $M_1 = D_2 + E_1$. Subsequently, by 3×3 convolutional operations, M_1, M_2, M_3, and M_4 are used to generate four feature maps denoted by P_1, P_2, P_3, and P_4. Finally, P_5 is generated by downsampling P_4. Five feature maps are resulted and denoted as $\{P_1, P_2, P_3, P_4, P_5\}$. These feature maps are then respectively inputted into RPN to generate region proposals and perform bounding boxes regression.

3.3. Customized anchor boxes

The concept of anchor boxes was initially introduced by (Ren et al., 2016). Anchor boxes can be defined as a set of bounding boxes with predefined scales and aspect ratios. These anchor boxes are evenly distributed across a feature map, strategically covering different positions. During the object detection process, each anchor box, is projected back onto the original image for comparison with the ground-truth bounding boxes, which define the true object locations. By establishing a set of anchor boxes with various scales and aspect ratios, an object detection model can gain flexibility in capturing objects of various sizes and shapes within the image. These anchor boxes serve as reference templates that provide spatial context to guide the subsequent detection process.

The intersection of union (IoU) between an anchor box and a ground-truth bounding box is used to estimate whether the anchor's position is

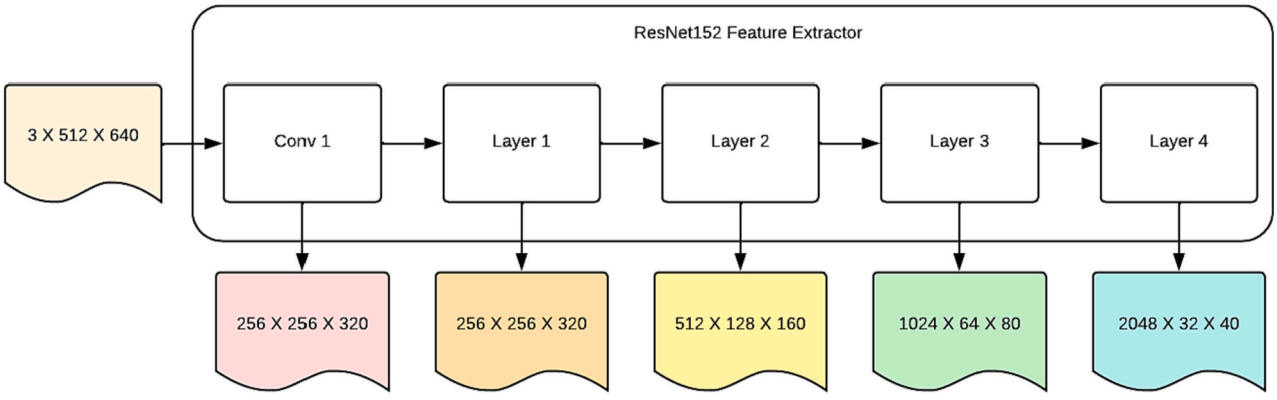


Fig. 4. The structure of ResNet152 Feature Extractor.

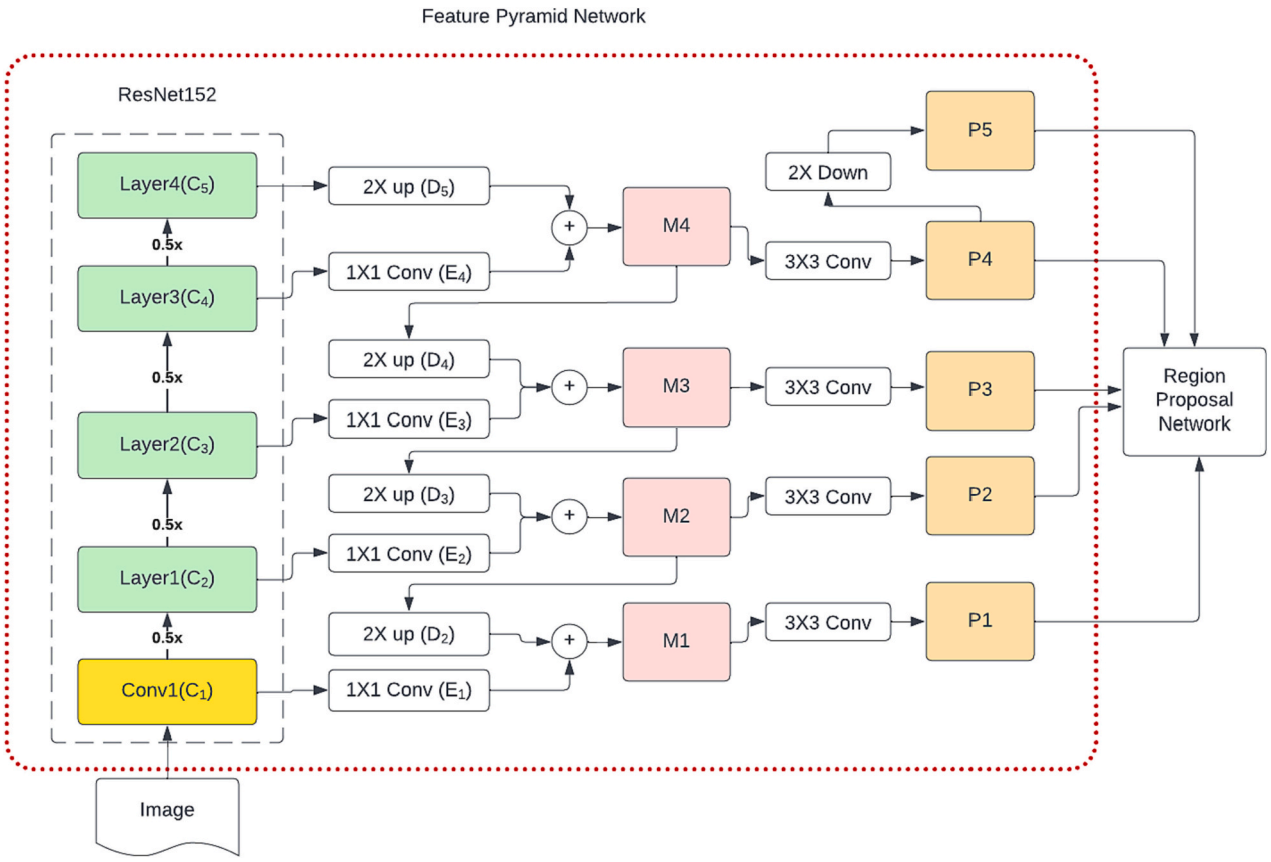


Fig. 5. The structure of FPN based on ResNet152.

the target's position. Suppose that A denotes an anchor box and B denotes a ground-truth bounding box, the algorithm of IoU is given by the following equation.

$$IoU = \frac{A \cap B}{A \cup B} \tag{1}$$

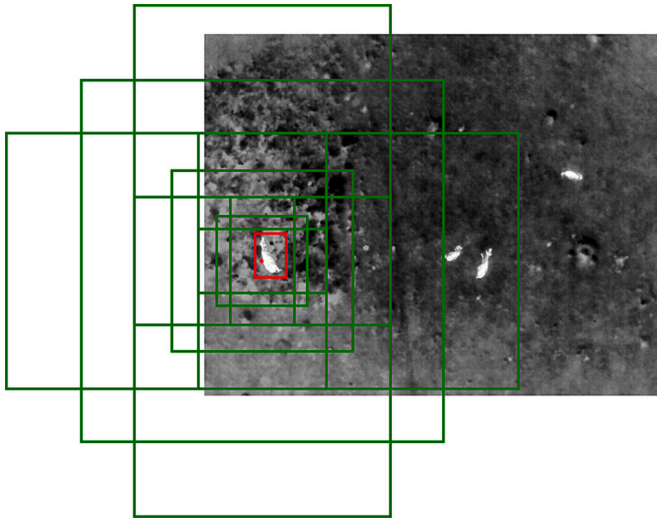


Fig. 6. An example of using anchors to detect deer. The red rectangle is a ground-truth bounding box, and the green rectangles are the anchors generated by three scales $(128^2, 256^2, 512^2)$ and three aspect ratios $(0.5, 1, 2.)$ (For interpretation of the references to colour in this figure legend the reader is referred to the web version of this article.)

The IoU between an anchor box and a ground-truth box serves as a primary measure of their proximity. Intuitively, with the increasing of the IoU, object A becomes more like object B. Generally, the IoU threshold is typically set to 0.5. When the IoU of an anchor box is <0.5, it is considered not to enclose the target. Conversely, if the IoU is equal to or >0.5, the anchor box is selected as a RoI for the subsequent bounding box regression. In original Faster R-CNN (Ren et al., 2016), three scales $(128^2, 256^2, 512^2)$ and three aspect ratios $(0.5, 1, 2)$ are used to yield 9 anchor boxes for each sliding. However, the animal objects in UAV thermal images are small and the average size is approximate to be 30 × 30 pixels, as shown in Fig. 6. When employing the anchor boxes defined in (Ren et al., 2016), the IoU values of anchor boxes tend to be <0.5. Therefore, these anchor boxes will be discarded, causing many small deer objects to be missed by the original Faster R-CNN. To address this problem, two tactics are employed in this paper. Firstly, the sizes of anchor boxes are systematically reduced. Secondly, the number of anchor boxes is increased at each position. Specifically, the anchor box scales are customized to be $(4^2, 8^2, 16^2, 32^2, 64^2)$, while keeping the aspect ratios the same. Therefore, at each position in a feature map, 15 different region proposals can be created. As shown in Fig. 7, the new strategies can ensure that at least one of the 15 region proposals intersects the ground-true bounding box and the corresponding IoU value is >0.5, enhancing the ability of the proposed model to capture and detect small objects from UAV thermal images.

Additionally, the bounding box regression in this paper is defined as the following.

$$t_x = \frac{G_x - A_x}{A_w}$$

$$t_y = \frac{G_y - A_y}{A_h}$$

$$t_w = log\left(\frac{G_w}{A_w}\right)$$

$$t_h = log\left(\frac{G_h}{A_h}\right) \quad (2)$$

(A_x, A_y) and (G_x, G_y) are the centers of a predicted anchor A and a ground-truth bounding box G respectively, and (A_w, A_h) and (G_w, G_h) denote the width and height of a RoI A and a ground-truth bounding box G respectively. Strictly speaking, the transformation equation (Eq. (2)) exhibits non-linearity. However, when a RoI is similar enough to its corresponding ground-true box, Eq. (2) can be treated as a linear regression. Conversely, when a RoI and its ground-true box are significantly different, Eq. (2) poses a complex non-linear regression problem, making it very difficult to align a RoI with a ground-truth bounding box accurately. For example, as shown in Fig. 7, the red rectangle represents a ground-true box including a deer object. The green rectangles denote the customized anchor boxes. Among these, there is at least one anchor box, and its IoU value with the red rectangle is >0.5, which means proximity in shape, location, and size to a ground-truth box. In this case, Eq. (2) can be conceptualized as bounding-box regression from an anchor box to a nearby ground-truth box. Based on t_x, t_y, t_w, and t_h, a loss function can be defined to facilitate the adjustment of an anchor box, aligning it with its corresponding ground-truth box (depicted as the red rectangle in Fig. 7). The adjustment is achieved through the process of backpropagation within a neural network. The loss function is instrumental in quantifying the disparity between the predicted and actual bounding box. By minimizing this disparity during training through backpropagation, the neural network learns to improve the accuracy of predicting anchor box positions. This iterative optimization process ensures that the model adapts its anchor boxes to closely match the ground-truth boxes, enhancing the overall precision of object detection in the training phase.

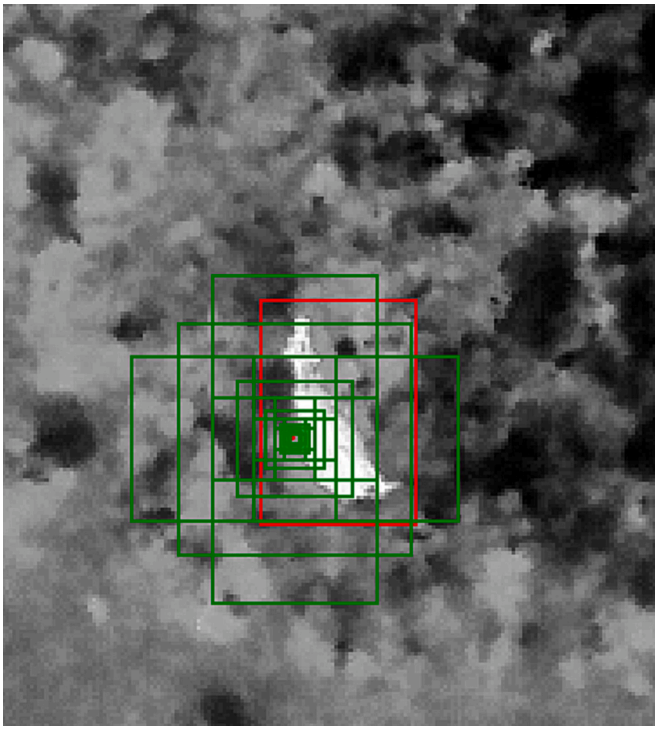


Fig. 7. An example of using 15 anchors to detect deer in this paper. The red rectangle is a ground-truth bounding box, and the green rectangles are the anchors generated by three scales $(4^2, 8^2, 16^2, 32^2, 64^2)$ and three aspect ratios $(0.5, 1, 2)$ (For interpretation of the references to colour in this figure legend the reader is referred to the web version of this article.)

3.4. Criterion for multi-scale feature map selection

In the original Faster R-CNN, only a single feature map is utilized, and all RoIs acquire their feature information from this common source. Thus, there is no ambiguity in feature map selection. However, utilizing a FPN built upon a residual neural network, it becomes possible to extract five feature maps with different resolutions from a thermal image captured by a UAV. Based on the small anchor boxes defined in Section 3.3, RPN can utilize these multi-scale feature maps to predict RoIs with various aspect ratios, leading to improved detection performance of small objects. However, when utilizing Fast R-CNN, as illustrated in Fig. 3 (b), for classifying these RoIs, a question arises: for a given RoI, which feature map should be selected to provide the feature information? Normally, the criterion is mainly based on the scale, size, and spatial characteristics of each RoI. In general, a larger RoI is assigned to a smaller-scale feature map, while a smaller RoI is assigned to a larger-scale feature map.

As shown on the top of Fig. 8, a feature map extractor based on FPN and ResNet152 extracted five feature maps from the original UAV thermal image, and they each has different resolutions, including 256 × 320, 128 × 160, 64 × 80, 32 × 40, and 16 × 20. The region proposal network can utilize the five feature maps to generate four RoIs with different sizes, which are respectively colored by purple, red, green and yellow from left to right in the middle-right of Fig. 8. The size of the purple RoI is the largest, and the model assigns a feature map with the resolution of 32 × 40 to it. On the contrary, the red RoI is the smallest, and the model selects the feature map with highest spatial resolution for it. Similarly, the green and yellow RoIs are assigned corresponding feature maps according to their respective sizes. The selection of large feature maps means the number of parameters is increasing and needs more time for training and detection, while the selection of small feature maps implies reducing the number of parameters and having faster detection speed. By this means, the model in this paper is allowed for a trade-off between detection accuracy and performance.

Based on the number of feature maps k, the RoI width w, and the RoI height h, a multi-scale feature map assignment criterion is expressed by the following equation:

$$index = floor\left(k + log_2\left(\sqrt{w \times h}\,/\,\lambda\right)\right)$$

$$if\ \sqrt{w \times h} \geq \lambda, \sqrt{w \times h} = \lambda \quad (3)$$

index is an integer value ranging from 1 to k, which denotes the index of the chosen feature map. For instance, the resolutions of the five feature maps on the top of Fig. 8 are 256 × 320, 128 × 160, 64 × 80, 32 × 40, and 16 × 20 respectively, and their indices are 1, 2, 3, 4, and 5 respectively. λ in Eq. (3) is a canonical adjustment parameter introduced in (Lin et al., 2016, 2018). When the square root of a RoI area is equal to or more than λ, it is regarded as a big object and assigned to the feature map with low spatial resolution. For example, the canonical pre-training size of the ImageNet Dataset is 224 pixels, therefore, most of the object detection models trained by the ImageNet dataset or true-color imagery usually set the value of λ to be 224. However, $\lambda = 224$ is not suitable for the detection of small objects in the UAV thermal images. λ is a empirically determined threshold. Therefore, λ in this paper was set to be 320 based on trials and errors.

4. Experiment

Based on different residual networks, five different FPNs were constructed, which are respectively named ResNet18FPN, ResNet34FPN, ResNet50FPN, ResNet101FPN, and ResNet152FPN. Each FPN serves as a backbone network to generate multi-scale feature maps. By the five FPNs, five different object detection models based on Faster R-CNN are

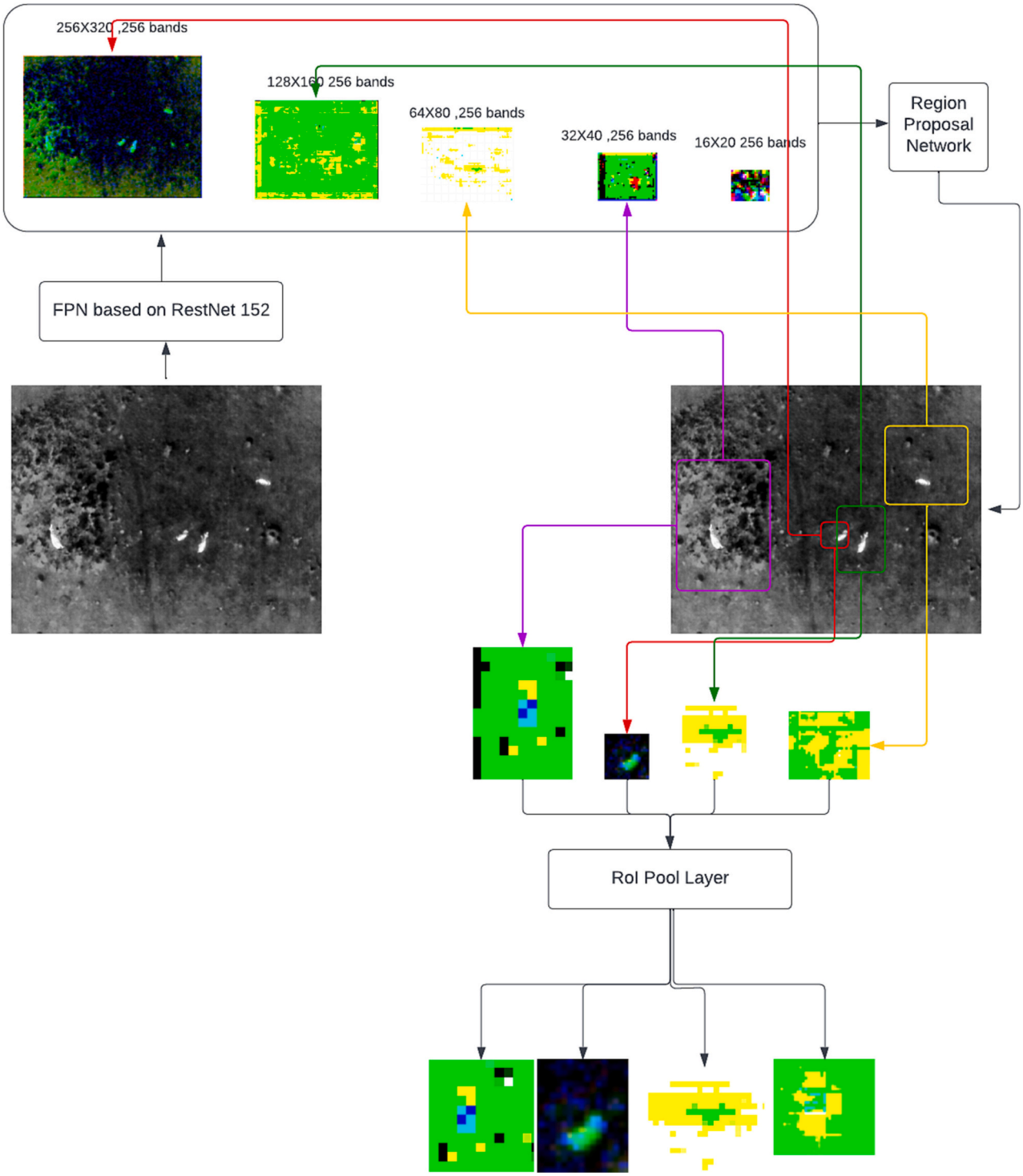


Fig. 8. The process that multiple-scale feature maps are assigned to the RoIs based on their resolutions.

constructed, and they are respectively named as FRC_ResNet18FPN, FRC_ResNet34FPN, FRC_ResNet50FPN, FRC_ResNet101FPN, and FRC_ResNet152FPN. In this section, we utilized the same thermal imagery dataset collected in the Chitwan National Park of Nepal to individually train each of the five object detection models. Subsequently, we conducted a comprehensive comparison of their performance of small deer object detection based on the COCO detection evaluation matrix (Padilla, Netto, and da Silva 2020). This evaluation aims to identify the most effective object detection model for wild deer surveys from UAV thermal images.

4.1. Data allocation

The wild deer survey in this paper was conducted in several conservation areas for wild deer in the Chitwan National Park of Nepal. The study areas are covered by riverine mixed forests and riparian grasslands, and the average temperature stays stable in a year ranging from 18 to 36C. Mavic 2 Enterprise Advanced DJI Drones, equipped with a thermal camera and a true-color camera fully stabilized by a 3-axis gimbal, were used to monitor wild deer. A total of 22,478 thermal images were captured, but not all these images contain wild deer. In order

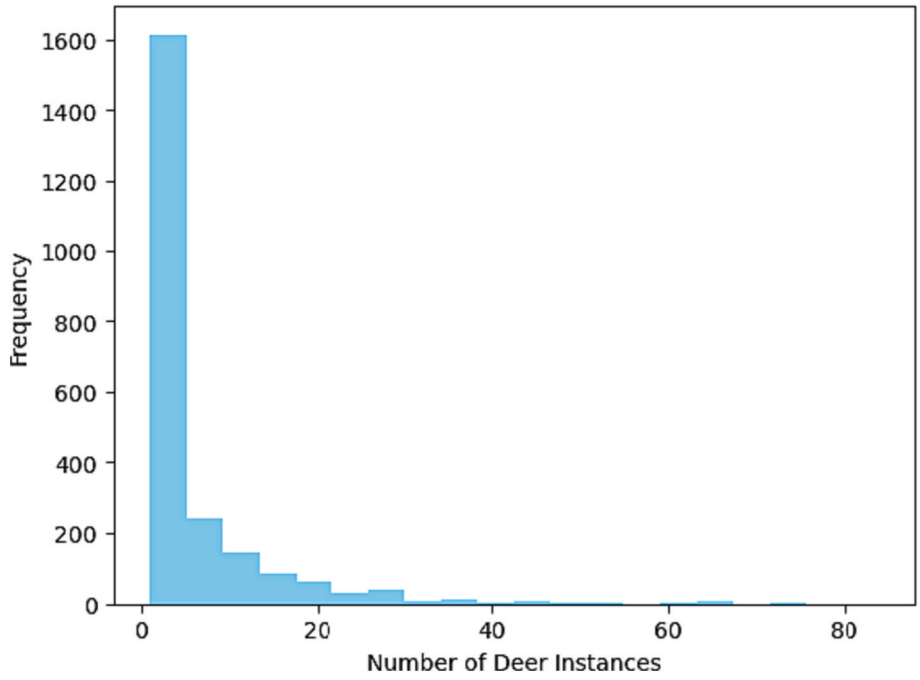


Fig. 9. The histogram of number of deer instances in the dataset. The number of bins is 20.

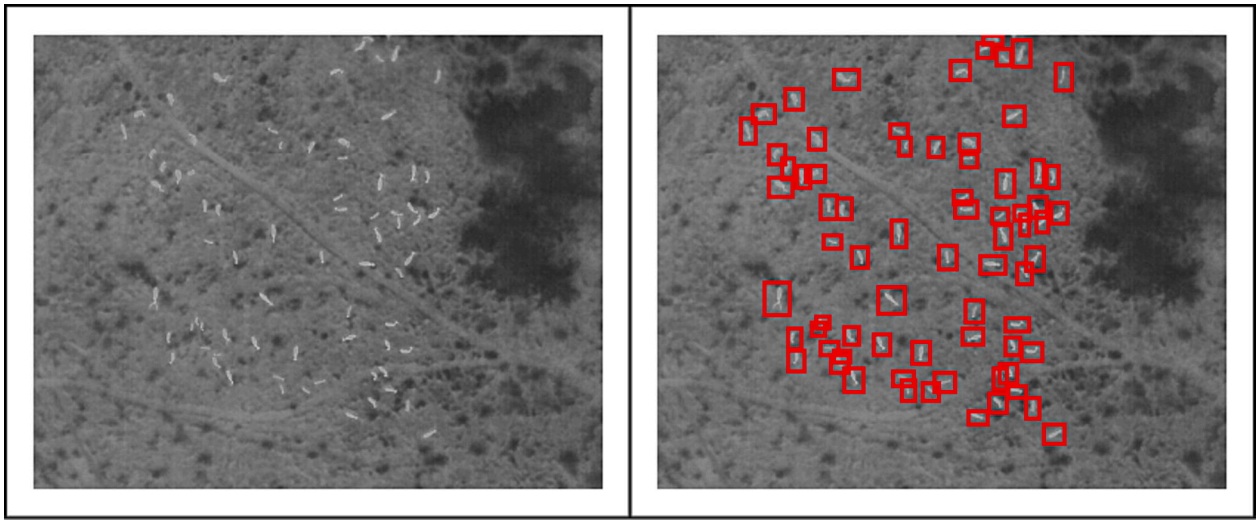


Fig. 10. An example of deer annotation. There are 72 deer in the left image. The 72 deer are annotated by ImageLab as shown in the right image.

to construct a UAV thermal image dataset for the training and validating the object detection models, 5000 thermal images that are likely to contain wild deer were manually soughted and each deer object was annotated by rectangles. Although this task was both time-consuming and labor-intensive, it is an indispensable process to build an automatic model for estimating the population of wild deer in the future. To ensure the precision of the annotations, two experienced scientists from the Center for Complex Human-Environment Systems in San Diego State University, along with two students from the University of Texas at Dallas, were employed. At first, the two scientists independently filtered the images by meticulously identifying the images including deer. Subsequently, the two students utilized 'ImageLab,' a freely available online image annotation tool, to label the precise locations of deer objects . Secondly, the two scientists then double checked all the annotations, cross-verifying against each other to determine whether an annotation should be considered as a ground truth or should be discarded. This verification process aims to prevent subjective biases and ensure the accuracy and reliability of the annotations. As a result, a comprehensive dataset consisting of 2278 thermal images and 13,509 deer instance annotations was constructed. The entire process took approximately seven days, with the two scientists and two students each dedicating four hours each day to complete the task.

The number of deer instances in each image varies, ranging from 1 to 72. The frequency histogram illustrating this distribution is depicted in Fig. 9, and an example image containing 72 deer annotations is displayed in Fig. 10. According to Fig. 9, it is evident that the distribution of deer numbers across the different images in the dataset is unbalanced, which may raise concerns if the dataset is randomly divided into training, validation, and testing subsets. To address this issue, the dataset is initially divided into separate subsets based on the number of deer in each image. Within each subset, the images are further partitioned into three sections using a ratio of 75:15:15. Subsequently, the

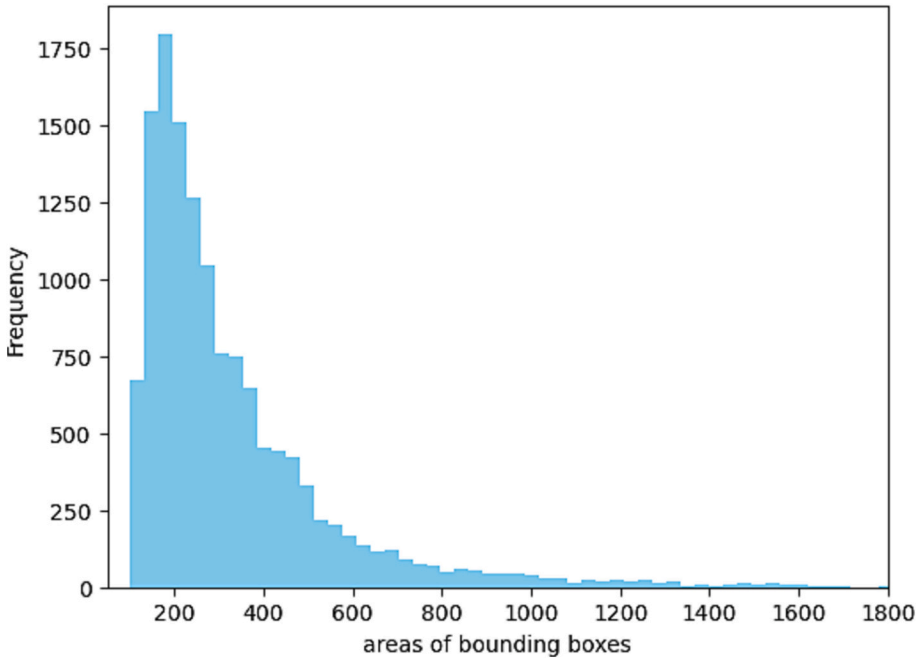


Fig. 11. The area histogram of the bounding boxes of deer objects in our dataset. The number of bins is 100.

images from different sections are combined to form three distinct datasets, which are respectively allocated for training, validation, and testing. By employing this approach, we ensure that each data set maintains a balanced distribution of deer instances.

4.2. Model evaluation criterion

To ensure a comprehensive assessment of the performance of models in this study, COCO detection evaluation matrix is adopted, and Average Precision (AP) and Average Recall (AR) are used to estimate the performance of different models (Padilla et al., 2020). AP and AR are computed based on the IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05, which accounts for varying levels of overlap between predicted and ground truth bounding boxes. AP provides a measure of the model's accuracy in positive predictions, focusing on the avoidance of false positives. It assesses how well the model's positive predictions align with the ground truth. On the other hand, AR measures the model's ability to correctly identify all positive instances, emphasizing the avoidance of false negatives. It evaluates how well the model captures all objects of interest. The AP and AR can offer insights into the model's precision and recall trade-offs, providing a method to assess the overall effectiveness of a model.

According to COCO detection evaluation matrix, all objects are categorized into three levels based on their sizes, including Large, Medium, and Small. The distribution of bounding box areas within the UAV image dataset in this paper is shown in Fig. 11. In fact, there is not a strict definition for Small Object, Medium Object, and Large Object. The criteria introduced in (Chen et al., 2017; Tong et al., 2020; Zhu et al., 2016) are widely accepted and are also adopted in this research. Based on the criteria, small objects are characterized by bounding boxes with an area range from 0 to 200 pixels. Medium objects are characterized by bounding boxes with an area range from 200 to 400 pixels. Large objects are characterized by bounding boxes with an area exceeding 400 pixels. The details on the number and ratio of Small Object, Medium Object, and Large Object in the UAV imagery in this paper are shown in Table 2. In Fig. 12, three example images are provided to show what small, medium, and large deer objects are like.

Table 2
The categorization criteria for deer objects based on their bounding boxes.

Level name	Range of area	Number	Ratio
Small Objects	100–200	3575	26.5%
Medium Objects	200–400	6721	49.8%
Large Objects	400–100,000	3213	23.7%

4.3. Experiment results analysis

By utilizing the five FPNs and small-scale anchor boxes $(4^2, 8^2, 16^2, 32^2, 64^2)$, we constructed five deer object detection models based on Faster R-CNN, namely FRC_ResNet18FPN, FRC_ResNet34FPN, FRC_ResNet50FPN, FRC_ResNet101FPN, and FRC_ResNet152FPN. For short, we also used M1, M2, M3, M4, and M5 to denote these five object detection models respectively in Table 3. The 13,509 deer objects were divided into the training set (9,115 objects), testing set (2,197 objects) and validation set (2197 objects). The average precision for small, medium, and large objects based on the test set as shown in Table 3.

$IoU \geq 0.5$ means that it will be a true positive prediction only when the IoU value between a predicted bounding box and its corresponding ground-true bounding box is greater than or equal to 0.5. $0.5 \leq IoU \leq 0.95$ means ten IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05, and ten precisions for different ranges can be calculated, and then compute the mean of the ten values. This comprehensive evaluation accounts for varying levels of overlap between predicted and ground truth bounding boxes, which is often used to evaluate how well the bounding boxes generated by models fit corresponding objects. According to the results in Table 3, the performance of both FRC_ResNet18FPN and FRC_ResNet152FPN is very close and surpasses that of the remaining three models. Under the condition of $IoU \geq 0.5$, the AP of the five models is shown in Fig. 13.

As shown in Table 3, the five models share a common characteristic. The Average Precision (AP) demonstrates a progressive decrease as the size of the objects being detected decreases. Generally, the detection of large deer objects tends to yield higher AP, as these objects are relatively easier to discern and locate accurately within the thermal images. As the size of deer objects decreases, the detection task becomes more challenging, leading to a decrease in the AP for both medium and small deer objects. In addition, AP is often sensitive to the IoU threshold, and strict

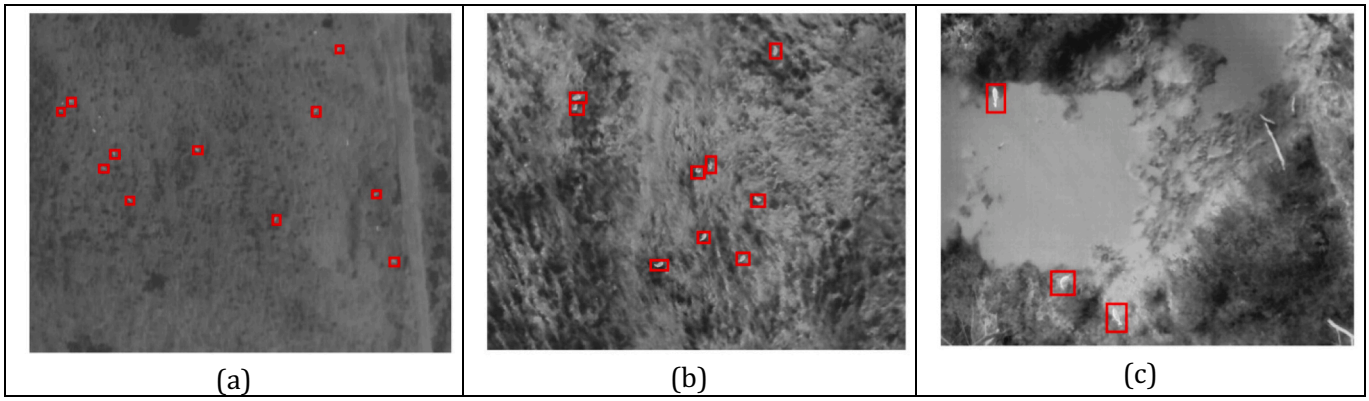


Fig. 12. (a) an example of small deer objects; (b) an example of medium deer objects; (c) an example of large deer objects.

Table 3
The COCO detection evaluation matrix of the five object detection models. M1 denotes FRC_ResNet18FPN. M2 denotes FRC_ResNet34FPN. M3 denotes FRC_ResNet50FPN. M4 denotes FRC_ResNet101FPN. M5 denotes FRC_ResNet152FPN.

	IoU	Object Size	M1	M2	M3	M4	M5
Average Precision	$IoU \geq 0.5$	All	91.6%	90.2%	90.4%	88.2%	90.4%
		Small	73.6%	70.6%	76.3%	68.4%	78.3%
		Medium	93.4%	90.3%	90.8%	88.6%	91.1%
		Large	94.3%	94.6%	92.6%	93.8%	92.2%
	$0.5 \leq IoU \leq 0.95$	All	44.1%	43.4%	44.4%	42.6%	44.2%
		Small	33.4%	31.2%	35.8%	30.4%	34.1%
		Medium	42.6%	38.7%	41.8%	38.2%	41.1%
		Large	47.1%	48.1%	47.6%	48.3%	48.5%
Average Recall	$0.5 \leq IoU \leq 0.95$	All	47.1%	48.2%	47.6%	48.3%	48.5%
		Small	50.4%	50.9%	50.4%	44.8%	46.9%
		Medium	53.1%	52.3%	53.8%	51.9%	53.7%
		Large	54.5%	54.4%	54.7%	54.4%	54.7%

criteria for overlap between predicted and ground true bounding boxes result in low AP values. Therefore, as the IoU threshold value increases, the AP of models also demonstrates a decreasing trend. Thus the AP under the condition of $0.5 \leq IoU \leq 0.95$ is less than that under the condition $IoU \geq 0.5$. According to Table 3, the FRC_ResNet18FPN achieves the best detection performance for medium objects (like in Fig. 12 (b)) with an AP of 93.4%, and the model FRC_ResNet34FPN has the best detection performance for large objects (like in Fig. 12 (c)) with an AP of 94.6%, which is only marginally higher by 0.3% compared to the FRC_ResNet18FPN, which obtains an AP of 94.3% for large objects. Generally, in deep convolutional neural networks (CNNs), the spatial resolution of feature maps typically diminishes with the increasing of layer depth. For small objects, their spatial features may become highly compressed, reducing to only a few pixels in deeper CNN layers. For instance, an object with dimensions of 15×15 pixels in a UAV thermal image might be represented by just 1 pixel in the feature map from Layer 4 of ResNet152 (refer to Fig. 4). The limited spatial resolution can lead to loss of fine details, making it difficult for the model to distinguish small objects from the background. Through FPN, different feature maps from different layers can complement each other, and deeper feature map can receive some spatial information from shallower layers. However, still certain spatial features might have been lost during the process of convolution operations. Consequently, the models of FRC_ResNet18FPN and FRC_ResNet34FPN can obtain higher AP for medium and large objects than the remaining models because they have less CNN layers than others. Notably, the FRC_ResNet152FPN obtains the best detection performance for small objects (Fig. 12 (a)) with an AP of 78.3%, which is explicitly higher than other models. Specifically, the AP of FRC_ResNet152FPN is approximately 5% higher than that of FRC_ResNet18FPN and 8% higher than that of FRC_ResNet34FPN. According to Table 2 and Fig. 12 (a), the sizes of small objects in the UAV imagery in this paper are $<$200 pixels, and the related spatial information is possibly not enough for a model to detect and identify them. Under this condition, more CNN layers mean that more abstractly semantic information can be extracted. Therefore, FRC_ResNet152FPN has more advantages than others in this sense.

For example, as shown in Fig. 14 (a), there are seven large deer objects in the UAV thermal image. The detection results of FRC_ResNet18FPN, FRC_ResNet34FPN, FRC_ResNet50FPN, FRC_ResNet101FPN, and FRC_ResNet152FPN are depicted in Fig. 14 (b), Fig. 14 (c), Fig. 14 (d), Fig. 14 (e), and Fig. 14 (f) respectively. The five models all successfully detect the seven deer object. However, FRC_ResNet18FPN and FRC_ResNet152FPN stand out by fitting the ground-truth bounding boxes more accurately compared to the other models. Certainly, there are also some UAV thermal images that the five models failed to detect all the objects. For example, as shown in Fig. 15, a thermal image has 17 deer objects. In this case, FRC_ResNet18FPN detects 15 deer objects, FRC_ResNet34FPN detects 11 deer objects, FRC_ResNet50FPN detects 12 deer objects, FRC_ResNet101FPN detects 14 deer objects, and FRC_ResNet152FPN detects 16 objects. In addition, as depicted by the yellow arrows in Fig. 15 a, the two close deer objects were not detected by all the five models.

In Fig. 16 (a), there are four small deer objects annotated by red rectangles, and there are also two deer objects that are not annotated intentionally, as indicated by two arrows. The detection results and corresponding IoU values generated by FRC_ResNet18FPN, FRC_ResNet34FPN, FRC_ResNet50FPN, FRC_ResNet101FPN, and FRC_ResNet152FPN are depicted in Fig. 16 (b), Fig. 16 (c), Fig. 16 (d), Fig. 16 (e), and Fig. 16 (f) respectively. All models can detect the four deer objects. FRC_ResNet18FPN and FRC_ResNet34FPN can also detect the two unlabeled deer objects. However, FRC_ResNet34FPN generates three extra false positive objects. From the perspective of IoU, the bounding boxes predicted by FRC_ResNet152FPN can fit the objects best. However, FRC_ResNet18FPN has better generalization ability than others in terms

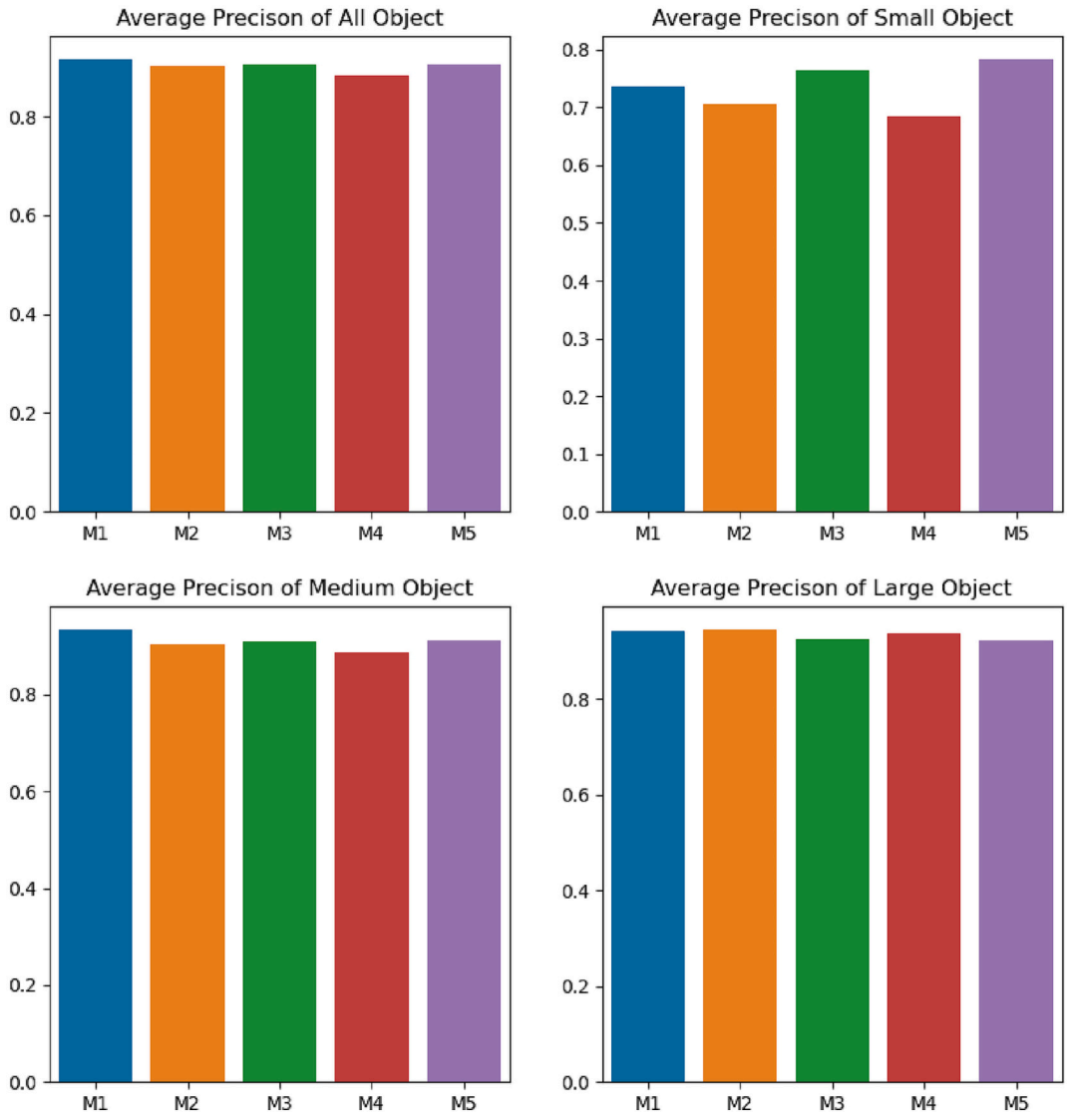


Fig. 13. The average precision of M1, M2, M3, M4, and M5. M1 denotes FRC_ResNet18FPN. M2 denotes FRC_ResNet34FPN. M3 denotes FRC_ResNet50FPN. M4 denotes FRC_ResNet101FPN. M5 denotes FRC_ResNet152FPN.

of identifying unseen objects.

In addition, we explore the impact of different scales of anchor boxes: our customized anchor boxes $(4^2, 8^2, 16^2, 32^2, 64^2)$ and the commonly used big-scale anchor boxes $(128^2, 256^2, 512^2)$. A thermal image containing 30 deer objects was used as an example. The model FRC_ResNet152FPN configured with big-scaled anchor boxes was able to detect 27 out of the 30 deer objects. However, three deer objects were overlooked, as highlighted by red arrows in Fig. 17 (a). On the contrary, the model FRC_ResNet152FPN equipped with the small-scaled anchor boxes exhibited a notably better outcome, successfully identifying and locating all 30 deer objects present in Fig. 17 (b). The experiment results show that using customized anchor boxes is helpful for the models to enhance their abilities of small object detection.

5. Discussion

5.1. Contributions

With the development of deep learning, research on object detection for wild deer surveys from high-resolution and true-color images has made remarkable progress. However, the research of detecting small deer objects in thermal imagery captured by UAVs remains relatively unexplored. Generally, in an object detection model based on deep learning, convolutional neural networks (CNNs) are adopted to extract feature maps from original images. The process of feature map extraction is through progressively downscaling. For the large deer objects, their abstract feature information can be kept in the final feature maps. However, for the small deer objects, their feature information often disappears during the process of feature map extraction, which is the main reason that small object detection is always challenging. In this paper, the integration of Faster R-CNN, FPN and residual networks is introduced to solve the problem of wild deer surveys from thermal images. To address the problem that the feature information of small deer objects disappearing during the process of feature map extraction, a Feature Pyramid Network (FPN) (Lin et al., 2016; Liu and Wang, 2021) is used to fuse the spatial feature information, derived from the different CNN layer of a residual network, to construct multiple feature maps for the detection of deer objects with different scales. At the same time, small-scaled anchor boxes were designed to serve as reference templates to provide more suitable spatial context to guide the detection process of small objects. Specifically, rather than employing commonly used large anchor boxes $(128^2, 256^2, 512^2)$, customized anchor boxes $(4^2, 8^2, 16^2, 32^2, 64^2)$ were utilized to generate the regions of interest

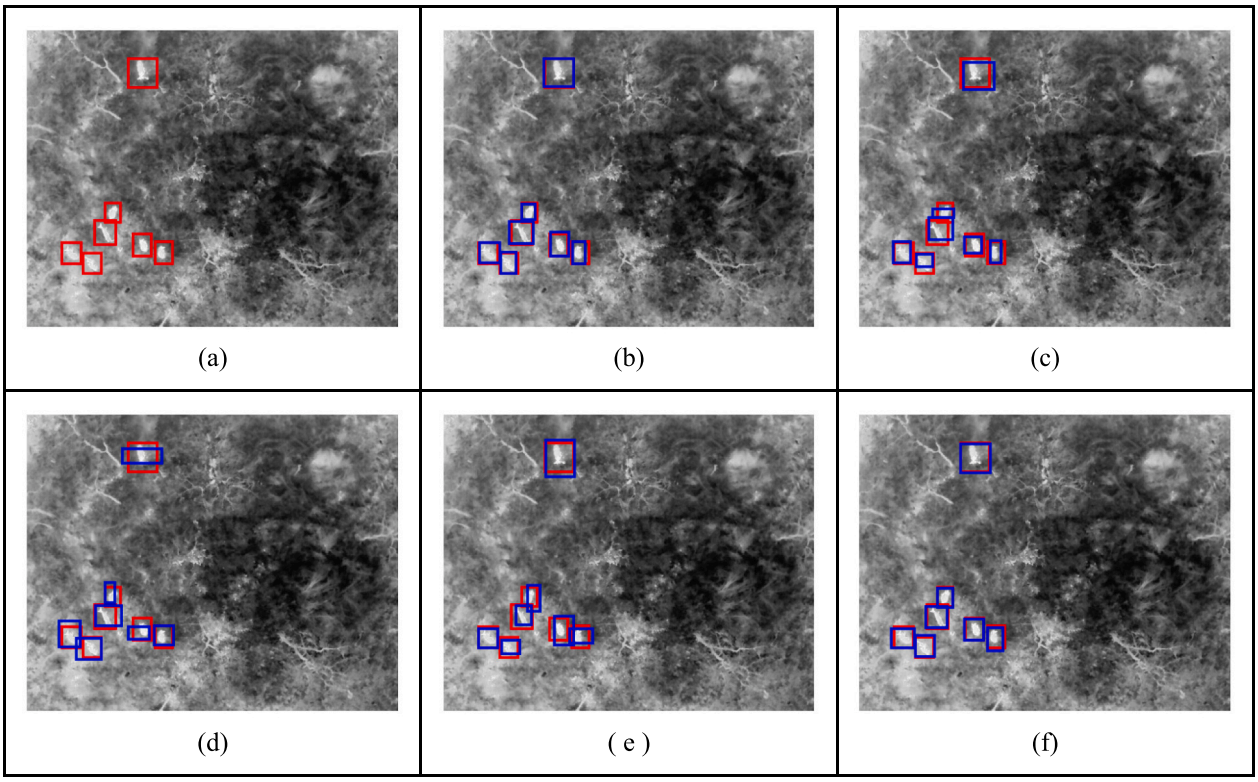


Fig. 14. (a) denotes the original image with the ground-true bounding boxes marked in red. (b) denotes the output of FRC_ResNet18FPN. (C) denotes the output of FRC_ResNet34FPN. (d) denotes the output of FRC_ResNet50FPN. (e) denotes the output of FRC_ResNet101FPN. (f) denotes the output of FRC_ResNet152FPN.

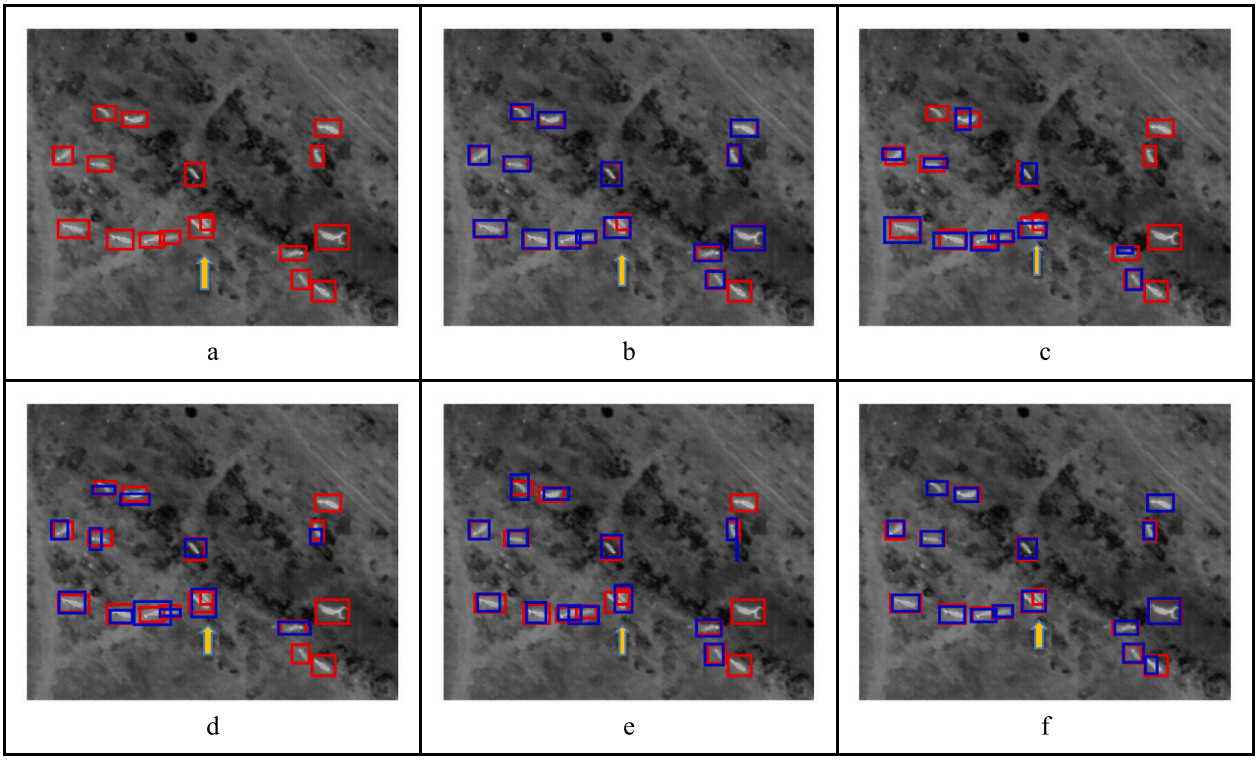


Fig. 15. (a) denotes the original image with the ground-true bounding boxes marked in red. (b) denotes the output of FRC_ResNet18FPN. (C) denotes the output of FRC_ResNet34FPN. (d) denotes the output of FRC_ResNet50FPN. (e) denotes the output of FRC_ResNet101FPN. (f) denotes the output of FRC_ResNet152FPN.

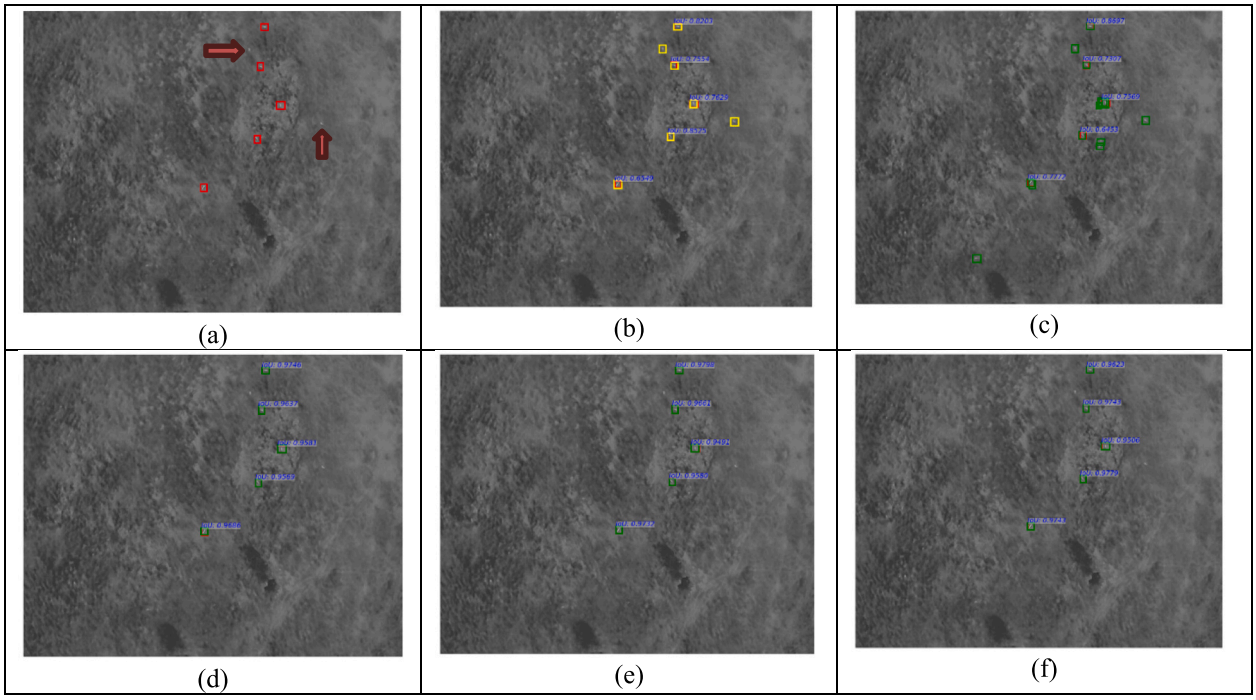


Fig. 16. (a) denotes the original image with the ground-true bounding boxes marked in red. (b) denotes the output of FRC_ResNet18FPN. (C) denotes the output of FRC_ResNet34FPN. (d) denotes the output of FRC_ResNet50FPN. (e) denotes the output of FRC_ResNet101FPN. (f) denotes the output of FRC_ResNet152FPN.

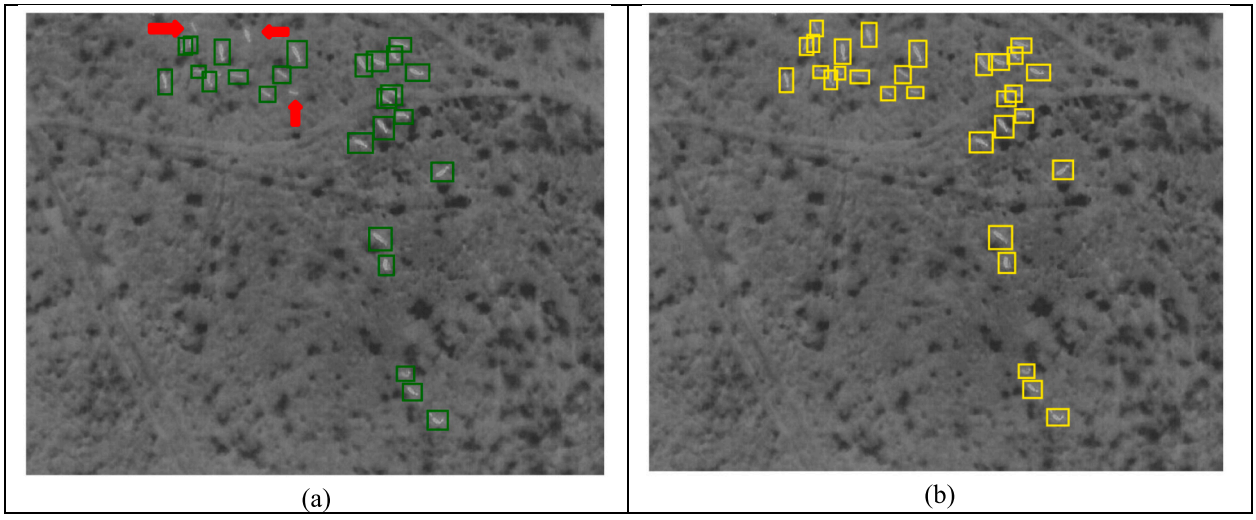


Fig. 17. (a) The detection results from the model using big-scale anchor boxes. (b) The detection results from the model using small-scale anchor boxes. The three red arrows point at the objects missed by the model in (a). (*For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.*)

(RoIs) based on feature maps at varying scales, which can enhance our model's capability to effectively detect small deer objects within thermal images. Finally, based on Faster R-CNN, FPN, and different residual networks including ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152 (Ganesan and Santhanam, 2022; He et al., 2016), we constructed five object detection models, and used the dataset to evaluate their detection performance by the COCO evaluation matrix. Our research endeavor is helpful for effective wild deer monitoring and conservation, providing valuable insights into deer populations in the Chitwan National Park of Nepal. The research outcomes can be a valuable reference for the exploration of small object detection from low-resolution thermal images.

5.2. Future work

Developing a detection model with the capability to handle various animal types in thermal imagery is a significant and promising direction for future research. Currently, our study focused only on deer detection due to limitations of data coverage. Multiple field research teams, supported by the same project, are exploring DJI drones images to also survey wild elephants, buffalos, rhinos, and wild boars in other areas of Chitwan National Park of Nepal. Wild animals in thermal imagery often exhibit relatively small sizes and similar shapes, which also pose additional challenges. Therefore, our future work is to construct an object detention model to automatically detect different small wildlife objects from thermal images and identify their species at the same time.

In addition, it is always important to explore methods to enlarge the feature information of a thermal image, especially for small animal objects detection and their species identification. DJI has released ten professional palettes designed to improve object features with varying temperatures in thermal images. Therefore, we hope to find some methods to transform a thermal image to multiple layers of images with different thermal palettes, and then to fuse feature information from these layers, which may further improve the performance of small animal objects detection.

Finally, as depicted by the yellow arrows in Fig. 15 (a), the two closely located deer objects were not detected by all the five models. For current object detection models, detecting close objects remains very challenging. This is due to the limitation of the Non-Maximum Suppression (NNS) algorithm. NMS relies on the Intersection over Union (IoU) threshold to determine whether two bounding boxes are considered duplicates or not. When objects are very close to each other, their bounding boxes may significantly overlap, leading to high IoU values. As a result, NMS may remove one of the objects, making it challenging for the model to detect both objects accurately. Therefore, new algorithms should be developed to solve the problem in the future.

6. Conclusion

Wild deer surveys are essential for wildlife management and conservation. By accurately estimating the number of deer, wildlife authorities and conservationists can better understand the health of the deer population, their interactions with other species, and their impact on the ecosystem. In the Chitwan National Park of Nepal, the dense coverage of tall trees and vegetation often obscures the presence of wild deer, making it very difficult to use normal true-color images to monitor deer. In our project, UAVs equipped with thermal cameras were used to monitor deer. However, thermal images have obvious limitations, such as lack of fine details, reduced spatial resolution, and limited spectral information. It is difficult to directly apply the traditional Faster R-CNN to detect deer objects in thermal images. In this paper, an enhanced Faster R-CNN based on FPN, residual networks and customized anchor boxes is proposed to detect small objects from UAV thermal images for wild deer survey in the Chitwan National Park of Nepal.

Specifically, based on Faster R-CNN, FPN, and different residual networks including ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152, five models are constructed. UAV thermal imagery including 2278 thermal images and 13,509 deer instance annotations were established to train, validate and test these models. At the same time, according to the sizes of deer objects, 13,509 deer instances are further divided into three categories: Small, Medium, and Large. Small objects are characterized by bounding boxes with an area range from 0 to 200 square pixels. Medium objects are characterized by bounding boxes with an area range from 200 to 400 square pixels. Large objects are characterized by bounding boxes with an area exceeding 400 square pixels. Finally, the COCO object estimation matrix was used to assess the performance of the five models. The COCO evaluation results revealed that under the condition of $IoU \geq 0.5$, the integration of Faster R-CNN, FPN, and ResNet18 is proved to be better than others, and achieved an Average Precision (AP) score of 91.6% for all deer objects. Specifically, the model obtained an AP score of 73.6% for small deer objects (area $\leq$ 200 pixels), an AP score of 93.4% for medium deer objects (200 $<$ area $\leq$ 400 pixels), and an AP score of 94.3% for large deer objects (area $>$ 400 pixels).

Funding

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

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

References

Aburasain, R.Y., Edirisinghe, E.A., Albatay, Ali, 2021. Drone-based cattle detection using deep neural networks. In: Arai, Kohei, Kapoor, Supriya, Bhatia, Rahul (Eds.), Intelligent Systems and Applications, Advances in Intelligent Systems and Computing. Springer International Publishing, Cham, pp. 598–611. https://doi.org/10.1007/978-3-030-55180-3_44.

Barbedo, Jayme Garcia, Arnal, Luciano Vieira, Koenigkan, Thiago Teixeira, Santos, and Patrícia Menezes Santos., 2019. A study on the detection of cattle in UAV images using deep learning. Sensors 19 (24), 5436. https://doi.org/10.3390/s19245436.

Barbedo, Jayme Garcia, Arnal, Luciano Vieira Koenigkan, Santos, Patrícia Menezes, 2020. Cattle detection using oblique UAV images. Drones 4 (4), 75. https://doi.org/10.3390/drones4040075.

Bengsen, Andrew J., Forsyth, David M., Pople, Anthony, Brennan, Michael, Amos, Matt, Leeson, Mal, Cox, Tarnya E., et al., 2022a. Effectiveness and Costs of Helicopter-Based Shooting of Deer. Wildlife Research.

Bengsen, Andrew J., Forsyth, David M., Ramsey, Dave S.L., Amos, Matt, Brennan, Michael, Pople, Anthony R., Comte, Sebastien, Crittle, Troy, 2022b. Estimating deer density and abundance using spatial mark–resight models with camera trap data. J. Mammal. 103 (3), 711–722. https://doi.org/10.1093/jmammal/gyac016.

Bharati, Puja, Pramanik, Ankita, 2020. Deep learning techniques—R-CNN to mask R-CNN: a survey. In: Das, Asit Kumar, Nayak, Janmenjoy, Naik, Bighnaraj, Pati, Soumen Kumar, Pelusi, Danilo (Eds.), Computational Intelligence in Pattern Recognition, Advances in Intelligent Systems and Computing. Springer, Singapore, pp. 657–668. https://doi.org/10.1007/978-981-13-9042-5_56.

Bochkovskiy, Alexey, Wang, Chien-Yao, Liao, Hong-Yuan Mark, 2020. YOLOv4: optimal speed and accuracy of object detection. arXiv. https://doi.org/10.48550/arXiv.2004.10934.

Carl, Christin, Schönfeld, Fiona, Profft, Ingolf, Klamm, Alisa, Landgraf, Dirk, 2020. Automated detection of European wild mammal species in camera trap images with an existing and pre-trained computer vision model. Eur. J. Wildl. Res. 66, 1–7.

Chen, Chenyi, Liu, Ming-Yu, Tuzel, Oncel, Xiao, Jianxiong, 2017. R-CNN for small object detection. In: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part V 13, 214–30. Springer.

Conner, Mary M., McKeever, Jane S., 2020. Are composition surveys for mule deer along roads or from helicopters biased? Lessons from the field. Wildl. Soc. Bull. 44 (1), 142–151.

Delplanque, Alexandre, Foucher, Samuel, Lejeune, Philippe, Linchant, Julie, Théau, Jérôme, 2022. Multispecies detection and identification of African mammals in aerial imagery using convolutional neural networks. Rem. Sens. Ecol. Conserv. 8 (2), 166–179. https://doi.org/10.1002/rse2.234.

Eikelboom, Jasper A.J., Wind, Johan, van de Ven, Eline, Kenana, Lekishon M., Schroder, Bradley, de Knegt, Henrik J., van Langevelde, Frank, Prins, Herbert H.T., 2019. Improving the precision and accuracy of animal population estimates with aerial image object detection. Methods Ecol. Evol. 10 (11), 1875–1887. https://doi.org/10.1111/2041-210X.13277.

Forsyth, David M., Comte, Sebastien, Davis, Naomi E., Bengsen, Andrew J., Côté, Steeve D., Hewitt, David G., Morellet, Nicolas, Mysterud, Atle, 2022. Methodology matters when estimating deer abundance: a global systematic review and recommendations for improvements. J. Wildl. Manag. 86 (4), e22207 https://doi.org/10.1002/jwmg.22207.

Freeman, Marianne S., Dick, Jaimie T.A., Reid, Neil, 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. Eco. Inform. 69, 101683.

Ganesan, Annalakshmi, Santhanam, Sakthivel Murugan, 2022. A novel feature descriptor based coral image classification using extreme learning machine with ameliorated chimp optimization algorithm. Eco. Inform. 68, 101527.

Girshick, Ross, 2015. Fast R-CNN. arXiv. https://doi.org/10.48550/arXiv.1504.08083.

Goyal, Vinat, Singh, Rishu, Dhawley, Mrudul, Kumar, Aveekal, Sharma, Sanjeev, 2023. Aerial object detection using deep learning: A review. In: Anupam Shukla, B.K., Murthy, Nitasha Hasteer, Van Belle, Jean-Paul (Eds.), Computational Intelligence. Springer Nature, Singapore, pp. 81–92. Lecture Notes in Electrical Engineering. https://doi.org/10.1007/978-981-19-7346-8_8.

Greenberg, Saul, Godin, Theresa, Whittington, Jesse, 2019. Design patterns for wildlife-related camera trap image analysis. Ecol. Evol. 9 (24), 13706–13730.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian, 2016. Deep Residual Learning for Image Recognition, pp. 770–778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Jobin, Benoît, Labrecque, Sandra, Grenier, Marcelle, Falardeau, Gilles, 2008. Object-based classification as an alternative approach to the traditional pixel-based classification to identify potential habitat of the grasshopper sparrow. Environ. Manag. 41, 20–31.

Kaur, Jaskirat, Singh, Williamjeet, 2022. Tools, techniques, datasets and application areas for object detection in an image: a review. Multimed. Tools Appl. 81 (27), 38297–38351.

Kellenberger, Benjamin, Marcos, Diego, Tuia, Devis, 2018. Detecting mammals in UAV images: best practices to address a substantially imbalanced dataset with deep learning. Remote Sens. Environ. 216 (October), 139–153. https://doi.org/10.1016/j.rse.2018.06.028.

Lautenschlager, R.A., 2021. Deer (Track-Pellet). In: CRC Handbook of Census Methods for Terrestrial Vertebrates. CRC Press, pp. 249–250.

Lin, Tsung-Yi, Dollár, Piotr, Girshick, Ross, He, Kaiming, Hariharan, Bharath, Belongie, Serge, 2016. Feature pyramid networks for object detection. arXiv.Org. (December 9, 2016) https://arxiv.org/abs/1612.03144v2.

Lin, Tsung-Yi, Goyal, Priya, Girshick, Ross, He, Kaiming, Dollár, Piotr, 2018. Focal loss for dense object detection. arXiv. https://doi.org/10.48550/arXiv.1708.02002.

Liu, Yong, Wang, Shengnan, 2021. A quantitative detection algorithm based on improved faster R-CNN for marine benthos. Eco. Inform. 61, 101228.

Nazir, Sajid, Kaleem, Muhammad, 2021. Advances in image acquisition and processing technologies transforming animal ecological studies. Eco. Inform. 61, 101212.

Norouzzadeh, Mohammad Sadegh, Nguyen, Anh, Kosmala, Margaret, Swanson, Alexandra, Palmer, Meredith S., Packer, Craig, Clune, Jeff, 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc. Natl. Acad. Sci. 115 (25), E5716–E5725.

Padilla, Rafael, Netto, Sergio L., Eduardo, A., Da Silva, 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.

Peng, Jinbang, Wang, Dongliang, Liao, Xiaohan, Shao, Quanqin, Sun, Zhigang, Yue, Huanyin, Ye, Huping, 2020. Wild animal survey using UAS imagery and deep learning: modified faster R-CNN for kiang detection in Tibetan plateau. ISPRS J. Photogramm. Remote Sens. 169 (November), 364–376. https://doi.org/10.1016/j.isprsjprs.2020.08.026.

Petso, Tinao, Jamisola Jr, Rodrigo S., Mpoeleng, Dimane, Bennitt, Emily, Mmereki, Wazha, 2021. Automatic animal identification from drone camera based on point pattern analysis of herd behaviour. Eco. Inform. 66, 101485.

Popek, Lukasz, Perz, Rafal, Galiński, Grzegorz, 2023. Comparison of different methods of animal detection and recognition on thermal camera images. Electronics 12 (2), 270.

Preston, Todd M., Wildhaber, Mark L., Green, Nicholas S., Albers, Janice L., Debenedetto, Geoffrey P., 2021. Enumerating white-tailed deer using unmanned aerial vehicles. Wildl. Soc. Bull. 45 (1), 97–108. https://doi.org/10.1002/wsb.1149.

Rangdal, Mukesh B., Hanchate, Dinesh B., 2014. Animal detection using histogram oriented gradient. Int. J. Recent Innov. Trends Comp. Commun. 2 (2), 178–183.

Ren, Shaoqing, He, Kaiming, Girshick, Ross, Sun, Jian, 2016. Faster R-CNN: towards real-time object detection with region proposal networks. arXiv. https://doi.org/10.48550/arXiv.1506.01497.

Rey, Nicolas, Volpi, Michele, Joost, Stéphane, Tuia, Devis, 2017. Detecting animals in African savanna with UAVs and the crowds. Remote Sens. Environ. 200 (October), 341–351. https://doi.org/10.1016/j.rse.2017.08.026.

Schwarz, Carl J., Seber, George A.F., 1999. Estimating animal abundance: review III. Stat. Sci. 14 (4), 427–456. https://doi.org/10.1214/ss/1009212521.

Simonyan, Karen, Zisserman, Andrew, 2015. Very deep convolutional networks for large-scale image recognition. arXiv. https://doi.org/10.48550/arXiv.1409.1556.

Tong, Kang, Yiquan, Wu, Zhou, Fei, 2020. Recent advances in small object detection based on deep learning: a review. Image Vis. Comput. 97 (May), 103910 https://doi.org/10.1016/j.imavis.2020.103910.

Torney, Colin J., Dobson, Andrew P., Borner, Felix, Lloyd-Jones, David J., Moyer, David, Maliti, Honori T., Mwita, Machoke, Fredrick, Howard, Borner, Markus, Grant, J., Hopcraft, C., 2016. Assessing rotation-invariant feature classification for automated wildebeest population counts. PLoS One 11 (5), e0156342.

Vecvanags, Alekss, Aktas, Kadir, Pavlovs, Ilja, Avots, Egils, Filipovs, Jevgenijs, Brauns, Agris, Done, Gundega, Jakovels, Dainis, Anbarjafari, Gholamreza, 2022. Ungulate detection and species classification from camera trap images using RetinaNet and faster R-CNN. Entropy 24 (3), 353.

Yudin, Dmitry, Sotnikov, Anton, Krishtopik, Andrey, 2019. Detection of big animals on images with road scenes using deep learning. In: 2019 International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI), 100–1003. IEEE.

Zhang, Haiyang, Chen, Changxi, 2020. Design of Sick Chicken Automatic Detection System Based on improved residual network. In: 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), vol. 1. IEEE, pp. 2480–2485.

Zhu, Zhe, Liang, Dun, Zhang, Songhai, Huang, Xiaolei, Li, Baoli, Shimin, Hu., 2016. Traffic-sign detection and classification in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2110–2118.