WildlifeMapper: Aerial Image Analysis for Multi-Species Detection and Identification

Satish Kumar[1*] Bowen Zhang[1*] Chandrakanth Gudavalli[1] Connor Levenson[1] Lacey Hughey[2]
Jared A. Stabach[2] Irene Amoke[3] Gordon Ojwang'[4] Joseph Mukeka[5] Stephen Mwiu[5]
Joseph Ogutu[6] Howard Frederick[7] B.S. Manjunath[1]
[1]University of California Santa Barbara [2]Smithsonian National Zoo and Conservation Biology Institute [3]Kenya Wildlife Trust, [4]University of Groningen, [5]Wildlife Research and Training Institute
[6]University of Hohenheim, [7]Tanzania Wildlife Research Institute
(satishkumar,bowen68,chandrakanth,clevenson,manj)@ucsb.edu,
(stabachj, hugheyl)@si.edu, irene.amoke@kenyawildlifetrust.org,
(gordonojwang, simbamangu)@gmail.com, (jmukeka, smwiu)@wrti.go.ke

Abstract

We introduce WildlifeMapper (WM), a flexible model designed to detect, locate, and identify multiple species in aerial imagery. It addresses the limitations of traditional, labor-intensive wildlife population assessments that are central to advancing environmental conservation efforts worldwide. While a number of methods exist to automate this process, they are often limited in their ability to generalize to different species or landscapes due to the dominance of homogeneous backgrounds and/or poorly captured local image structures. WM introduces two novel modules that help to capture the local structure and context of objects of interest to accurately localize and identify them, achieving a state-of-the-art (SOTA) detection rate of 0.56 mAP. Further, we introduce a large aerial imagery dataset with more than 11k Images and 28k annotations verified by domain experts. WM also achieves SOTA performance on 3 other publicly available aerial survey datasets collected across 4 different countries, improving mAP by 42%. Source code and trained models are available at Github [1].

1. Introduction

This paper introduces WildlifeMapper (WM) - an automated and scalable method for counting wildlife in aerial imagery. Aerial wildlife surveys are recognized as a cornerstone of modern conservation biology. By facilitating large-scale biological monitoring in remote landscapes, this technique has underpinned the ability to track changes in the abundance and distribution of wildlife across open landscapes for decades. However, traditional survey approaches often rely on manual observers to identify, count, and validate species of interest. This labor-intensive process can be time-consuming and error-prone, with potential to limit the utility of final results [27, 35, 38].

Automated approaches offer a promising alternative for efficient and accurate detection of wildlife in aerial survey images. Recent work, for example, illustrates how artificial intelligence has been used to count a variety of species from the air, including antelope in grasslands [44], whales in the ocean [11], and seals on the beach [42]. When combined with advancements in low-cost, high-resolution imaging platforms (e.g., UAVs), these case studies underscore the potential for such data to significantly reduce the effort and cost of traditional wildlife census methods. However, the majority of these techniques struggle to generalize to new species or landscapes due to the dominance of homogeneous backgrounds and poorly captured local structures [7, 12, 18, 39].

WildlifeMapper overcomes these limitations by adapting a novel application of the segment anything transformer model [21]. This model combines high frequency component correlations and spatial correlations in the image data to generate a map of potential locations of objects of interest (i.e., wildlife, livestock). In addition, we address the challenge of identifying multiple species from a relatively small footprint in these images.

To demonstrate the WM analysis workflow, we provide a case study example across the Masai Mara Ecosystem in southwestern Kenya. Renowned for its rich biological diversity, the abundance of large mammals (such as buffalo (*Syncerus caffer*), giraffe (*Giraffa tippelskirchi*), and wildebeest (*Connochaetes taurinus*)) have declined precipitously over the past few decades [34]. Importantly, our analy-

[1] *https://github.com/UCSB-VRL/WildlifeMapper*
* refers to equal contribution

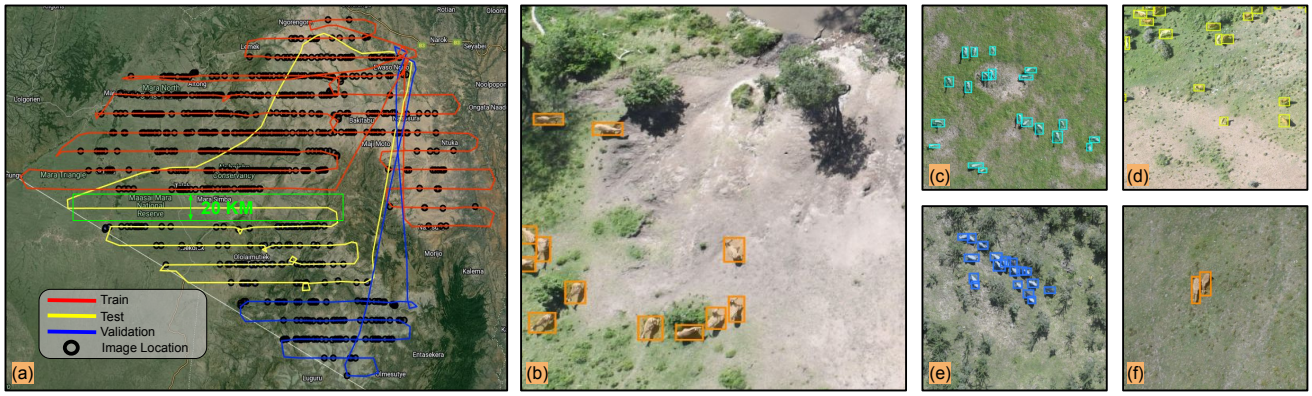

Figure 1. *Summary of Mara-Wildlife dataset. (a) Satellite view indicating the four flight trajectories, each represented in a different color. (b, c, d, e, f, g) Annotations of (b) zebra, (c) hartebeest, (d) cattle, (e) shoats (sheep and goats), and (f) zebra. Best viewed in color.*

sis incorporated $11,151$ images of size 8400×5500 collected from a digital camera affixed to the bellyport of a Partenavia P68 airplane during Systematic Reconnaissance Flight (SRF) surveys. Part of these images were annotated by trained observers with $28,146$ annotations of 21 species of large mammals ($\geq 15kg$), providing an unprecedented opportunity to develop species detection models across a complex, heterogeneous environment. The dataset was systematically verified by trained observers as described in Section 3. Our contributions can be summarized as follows:

1. A novel, single-stage end-to-end approach for animal detection. The modules, a *High Frequency Feature Generator*, a *Feature Refiner*, and a *Query Refiner*, work together to improve upon the traditional methods of object detection in aerial imagery and enable generalizability across different habitats. The high frequency features reduce dependence on dominant backgrounds/landscapes.

2. An input patch embedding layer that is specifically designed to capture contextual information to help in identifying individual animal species.

3. The release of a new benchmark dataset via the data owner (Kenya's Wildlife Research and Training Institute - WRTI) once all approvals are in place. An international user community is already engaged in further enhancing these data and using the WM through the BisQue [25] platform.

Community Adoption The practical use of WildlifeMapper (WM) extends beyond theoretical and computational success and is operationalized through BisQue [25]. WM is made available to users through a series of training modules that demonstrate how to *(i)* upload digital imagery, *(ii)* create annotations, *(iii)* apply and/or improve existing models, *(iv)* evaluate model fit, *(v)* improve annotations and refit models, and *(vi)* generate summary statistics. In the future, we envision WM to be of great value to ecologists, wildlife managers, and government officials, providing accurate information about the state of wildlife populations in near real-time, facilitating decision-making processes, and improving the conservation of ecosystems globally.

2. Related Works

Manual Methods: Aerial surveys using Front- and Rear-Seat Observers (FSO and RSO, respectively) are commonly used to inventory wildlife populations across open landscapes [36]. However, several important biases can impact these counts, including the experience level and fatigue of the human observers [28].

Deep Learning Methods: To address these issues, researchers have begun incorporating digital cameras on piloted aircraft and UAVs [18, 30, 46]. This minimizes the influence of observer bias while increasing transparency and reproducibility of results. For example, [26] replaced RSOs with an oblique camera mount system minimized the influence of observer bias while producing comparable estimates of large mammals under partial canopy cover. Similarly, [44] used a nadir mounted camera to improve the accuracy and efficiency of manual counts of large antelope in an open grassland ecosystem.

However, detecting animals in the wild from aerial imagery poses many challenges. For example, most publicly available datasets for aerial object detection are focused on identifying relatively distinct features such as buildings, roads, vehicles, and other man-made structures [8, 14, 53]. Animals, tend to blend in with their surroundings [13], can be occluded by trees, exhibit considerable variation in color and pattern, or have behavioral adaptations that make them difficult to detect [7, 20, 39, 48].

[20] proposed a solution for this problem involving a two-branch CNN model based on AlexNet to perform animal recognition and localization. [7] evaluated three state-of-the-art object detection algorithms, including Faster-RCNN, Libra-RCNN, and RetinaNet on six African wild

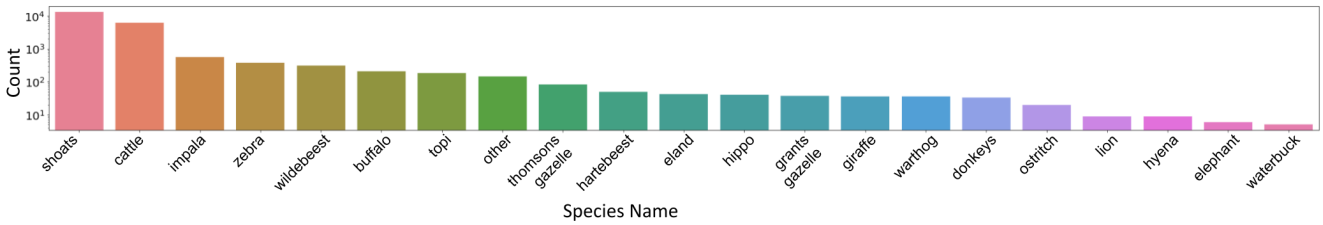

Figure 2. *Distribution of ($\geq$ 15kg) mammals identified in digital imagery collected across the Masai Mara Ecosystem, Kenya.*

Dataset	*# of annot. images*	*# of annot. tiles*	*# of species*	*# of annotation*	*Image size*	*GSD (cm)*	*Location*
Virunga	739	30069	6	5664	6000x4000	2.4	*DRC*
Garamba	158	6429	6	1611	6000x4000	2.0	*DRC*
AED	2067	69387	1	15581	5500x3600	2.4-13.0	*Bostwana, Namibia, South Africa*
Mara-Wildlife	**1012**	**77966**	**21**	**28146**	**8256x5504**	**1.45**	***Masai Mara National Reserve***

Table 1. *Comparison of Mara-Wildlife dataset with other publicly available dataset. Mara-Wildlife dataset has $\times$3 more unique species than the total of all other datasets. Each image is significantly larger and higher ground resolution making 77k unique images of size 1024 $\times$ 1024 with 21 different animal species. GSD: ground sampling distance; DRC: Democratic Republic of Congo.*

mammals. All three algorithms, however, showed poor performance in animal detection when animals were grouped closely together in herds. [12] adopted a segmentation approach, employing a UNet model to detect livestock from drone imagery. [48] uses a comparable model to analyze high-resolution satellite imagery, producing segmentation masks of wildebeest-sized animals, which are subsequently utilized for detection and counting.

Transformers: WM adopts a transformer architecture based on past success in modeling different types of aerial imagery. Examples include incorporating multispectral imagery for change detection [4], landcover classification [16], greenhouse gas (GHG) emission detection [23,24] and RGB aerial imagery for object detection [29, 49, 52].

The most effective applications of transformer-based models have been tailored for standard object detection tasks [3, 5, 32, 41, 54]. These works leverage the self-attention to model dependencies among the patches in an end-to-end fashion, unlike CNN-based models [1, 40]. However, when directly applied to aerial imagery, these models cannot effectively exploit the local structures as they divide the image into a sequence of patches. This limits the detection of small-scale objects in a homogeneous and dominant background.

Dataset: The existing publicly available animal aerial imagery datasets are listed in Table 1. The Virunga dataset [7] was collected in Virunga National Park, Democratic Republic of Congo (DRC). This dataset contains 897 annotated images of 6 animal species. The Garamba dataset was collected in Garamba National Park, DRC and contains 7034 images. Only 158 images have been annotated, containing 7 animal species. The aerial elephant dataset (AED) [33] was collected across a mosaic of woodland, open shrubland, and grassland habitats in Botswana, Namibia, and South Africa. Only a single species (i.e., elephant) was targeted during SRF surveys. See Table 1 for details.

3. Mara-Wildlife Dataset

The Mara-Wildlife dataset is a distinctive dataset that captures the essence of the Masai Mara ecosystem through a compilation of 77966 images of size 1024 x 1024. This habitat is heterogeneous, including woodland, shrubland, and grassland vegetation with 21 unique animal species.

3.1. Image Collection

Flightline Details: Data collection was in collaboration with the Smithsonian National Zoo and Conservation Biology Institute (SNZCBI), Kenya's Wildlife Research and Training Institute (WRTI), the Kenya Wildlife Trust (KWT), and the Directorate of Resource Surveys and Remote Sensing (DRSRS). In March 2022, we fitted a Partenavia P68 with a Nikon D850 digital camera and collected high resolution (8256 $\times$ 5504) digital images. During data acquisition, the aircraft adhered to a predetermined flight trajectory, depicted in Figure 1, at 400 ft above ground level (agl). This trajectory was optimized to encompass open grassland areas across the Masai Mara ecosystem, including the Masai Mara National Reserve, 22 adjacent private conservancies, and unprotected peripheral areas.

The aerial survey was conducted during a wet season period when the Serengeti migratory population of wildebeest have moved southward to locate more suitable forage in Tanzania. Thus, the survey primarily captured resident species, including wildebeest, zebra, topi, hartebeest, gi-

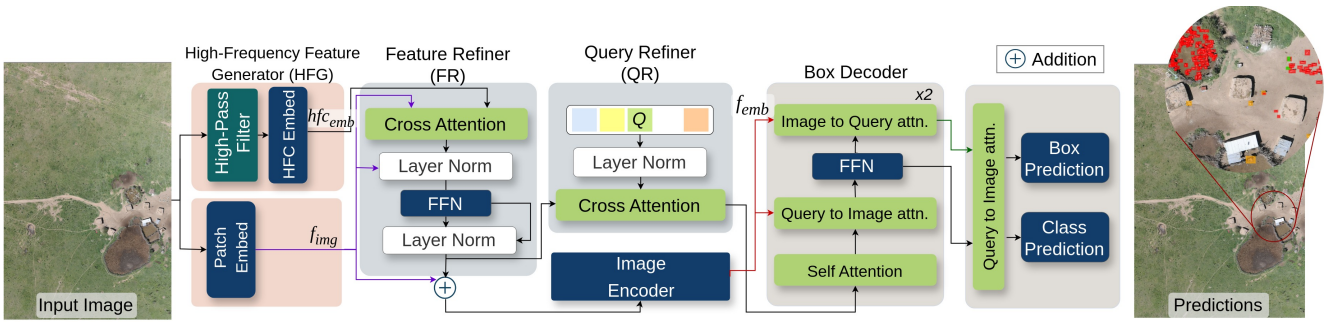

Figure 3. *Overview of WildlifeMapper (WM) architecture. Given an input image of size* $1024 \times 1024 \times 3$*, the High-Frequency Feature Generator (HFG) module generates information about potential location of object of interest. The Feature Refiner (FR) takes these potential location along with contextual features from Patch Embed layer and sent output to Image Encoder. In parallel, the Query Refiner (QR) incorporates the output of FR to refine learnable queries. Finally these queries are decoded using encoded features from Image Encoder and predict bounding box and class.*

raffe, and other large ($\geq$ 15 kg) antelope. Data collection was conducted in the early mornings (prior to 10:00 EAT) and late afternoons (after 15:00 EAT) when lighting conditions were optimal and animals are most active.

The geographical positioning of each image was acquired through a GPS system that recorded the plane's altitude, speed, and geographical coordinates. These data were then synchronized with the image using the image capture timestamps, enabling us to determine the geographic location of the centroid of each image. The camera was placed in the bellyport of the airplane, capturing a nadir view of the landscape every two seconds along the flight path.

3.2. Image Annotation

Initial bounding box annotations (21796) generated using AIDE platform [19] were exported in CSV format. These were then imported into BisQue [25] for further validation and correction. Finally, these annotations underwent validation by a single trained observer specializing in ecology, resulting in 28146 annotations in total.

3.3. Dataset Statistics

The Mara-Wildlife dataset showcases a detailed assemblage of wildlife, inclusive of 21 distinct species classes. The dataset is composed of approximately $77,966$ tiled images, derived from $1,012$ original rasters (Table. 1) The meticulous process of annotation has culminated in labeling $28,146$ animals. Species identified include domestic cattle (*Bos taurus*), white-bearded wildebeest (*Connochaetes taurinus*), topi (*Damaliscus lunatus*), shoats (domesticated sheep and goats), kongoni (*Alcelaphus buselaphus*), waterbuck (*Kobus ellipsiprymnus*), impala (*Aepyceros melampus*), Grant's gazelle (*Nanger granti*), Thomson's gazelle (*Eudorcas thomsonii*), Cape buffalo (*Syncerus caffer*), zebra (*Equus quagga*), ostrich (*Struthio camelus*), Masai giraffe (*Giraffa tippelskirchi*), warthog (*Phacochoerus africanus*), eland (*Taurotragus oryx*), donkey (*Equus africanus*), hyena (*Crocuta crocuta*), hippopotamus (*Hippopotamus amphibius*), lion (*Panthera leo*), and elephant (*Loxodonta africana*).

4. WildlifeMapper Architecture

4.1. Technical Overview

WildlifeMapper's architecture is inspired by the success of the Segment Anything Model (SAM) [21], created to segment small/large (all sizes) of objects. Referring to Fig. 3, WM contains the following main components: (i) A patch embedding layer designed to capture long-range context, (ii) a *High-Frequency Feature Generator* (**HFG**), (iii) a *Feature Refiner* (**FR**) followed by ViT based image encoder [21,23,45], and (iv) a *Query Refiner* (**QR**) module followed by a box decoder module. The input image is first processed through two separate branches, the patch embedding layer which captures long-range context [9] and **HFG** which suppresses all the low-frequency components in the image and generates a feature embedding. The **HFG** (Sec. 4.3) exploits prior knowledge that aerial images from areas such as forests, grasslands, and shrublands have a homogeneous and dominant background representing the dominant low frequency image content. Fig. 4 shows that on suppressing the lower frequencies, the object of interest is easy to locate.

The **FR** (Sec. 4.4) takes the embeddings from each of the two branches and generates a high quality embedding that contains information about potential locations of animals and captures the local context. The **QR** modules refines a set of learnable queries using the location information from **FR** module. These refined queries are passed to the box decoder. The box decoder takes the refined queries and encoded features image encoder to generate the final detection box and class of the object.

4.2. Patch Embed

The patch embeddidng layer utilizes a larger kernel convolution with an increasing dilation rate. This design rapidly expands the receptive field, allowing explicit extrac-

tion of features rich in contextual information. This approach is particularly beneficial for aerial imagery, where the small sized object makes classification based on appearance alone challenging. Contextual information thus becomes crucial for the accurate recognition of these objects.

4.3. High-frequency Feature Generator (HFG)

Along with patch embedding, the input image is processed in parallel by the **HFG** module to generate features with information about the location of the animal or cluster as shown in Fig. 4. The **HFG** module is inspired from the limitation of ViT models [45]. ViT models face challenges in efficiently utilizing local structures. They segment an image into patches and apply self-attention to model relationships, but this approach often falls short in capturing detailed local features [37, 51].

Local features in images are closely linked to high-frequency components [2, 43]. We hypothesize that suppressing low-frequency components can mitigate the influence of a dominant homogeneous background. To test this, we performed a discrete Fourier Transform (DFT) on the images, filtering out the low-frequency components before reconstructing the images, as shown in Fig. 4.

For a given input image $I \in \mathbb{R}^{H \times W \times C}$, where C is channel dimension, we compute Discrete Fourier Transform (DFT) of I. In next step we suppress the low frequency components with a controlling parameter and construct the image I with inverse DFT ($IDFT$).

$$I' = IDFT[hpf(DFT(I)] \tag{1}$$

where hpf is a high pass filter. Then we reduce the dimension of the reconstructed image I' via an embedding layer to generate embedding hfc_{emb} and pass them to the **FR** module. See supplementary materials for more details.

4.4. Feature Refiner (FR)

Next, the features from the patch embedding layer (f_{img}) and **HFG** (hfc_{emb}) are fed to the **FR** module. The f_{img} are refined with the hfc_{emb} via cross-attention mechanism. The **FR** is a simple module with cross-attention and linear layers [45]. The output contains information about the potential location of the object and the long-range context.

Following the standard architecture of SAM [21], we pass **FR** output to our ViT based image encoder supplemented with learnable positional embeddings p. The encoded feature map is f_{emb}:

$$f_{emb} = \text{ViT}\ [\mathbf{FR}(f_{img},\ hfc_{emb}),\ \ p] \tag{2}$$

Query Refiner (QR): The **QR** follows a transformer decoder like architecture and takes as input a set of 100 learnable queries $Q \in \mathbb{R}^{100 \times d}$ and output of **FR** module. Here $d = 256$ is same as channel dimension of f_{emb} from image encoder. The **FR** output refines the Q via a cross-attention

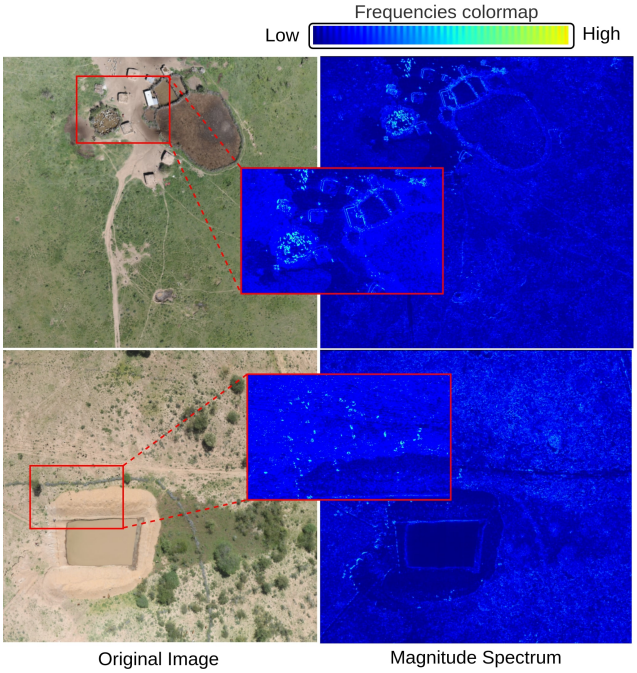

Figure 4. *The sample output visualization from the High-Frequency Feature Generator (**HFG**) module. The illustration shows the effectiveness of the module in suppressing the homogeneous and dominant background, while highlighting objects of interest (i.e., animals). The top image shows bomas, natural structures constructed to contain livestock, and paths that have been suppressed. Animals, however, are clearly identified, especially inside the boma. The bottom image shows a water body (a dam created for livestock) that has been suppressed by the module. Animals can again be highlighted throughout the image.*

mechanism. The refined queries narrows the search space for box decoder module to accurately locate and identify object of interest. [15, 23].

4.5. Box Decoder

Next, the refined queries are sent to the box decoder module. We concatenate these queries with the f_{emb} and pass box decoder's self attention layer as qf_{emb}. This is inspired from the idea of *class_token* used by Vaswani et.al. [45] to make decoding process memory efficient. The transformer network takes 4 input variables, those are *position embeddings of queries and image embeddings*, qf_{emb}, and f_{emb}. Our transformer model uses two-way attention inspired from [3, 6, 21] and our box decoder uses self-attention and cross-attention in two directions (queries-to-image embedding and vice-versa) to update all embeddings. We keep the box decoder very light weight (two blocks). The top 100 (equal to number of queries) indexes from the output of the final block is passed to two separate MLP blocks to regress the output bounding box prediction and class of the predicted box.

4.6. Training and Inference

We've trained WM using a single-stage, end-to-end approach to determine bounding boxes and classify them. The loss strategy we applied for WM is akin to what's used in DETR [3]. Initially, we perform bipartite matching to align our model's predictions with the actual bounding box data. Then, we proceed to compute the loss for these matched pairs. To achieve the best possible match between our predictions and the real data, we use the Hungarian algorithm [22]. Once matched, each prediction is paired with its respective ground truth. We then measure the l_1 (L1 distance) and $GIoU$ loss for the bounding box and the cross entropy loss for the classification [3].

Inference: The inference pipeline is straightforward and similar to training code. During inference, we first filter the detections at 50% threshold and then use non-maximal suppression to remove any overlapping boxes.

5. Experiments

Train-val-test split: There is no data leak in train-val-test split. Fig. 1 shows the flight paths with location of each image represented by a circle. Each color coded flight path represents the train-val-test image set. No images taken while the airplane was cutting across the transects plus turnaround points as can be seen there are no circles. We created a spatial disjoint of $20km$ distance between the transects as shown in Fig. 1 and achieved consistent performance ($0.56mAP$).

Evaluation Metrics: We report our performance on multiple metrics. Following the protocols of standard object detection, we report the performance in mean average-precision (mAP) for detection and mean intersection-over-union (mIOU) for localization of animals. We also report a commonly used metric by the ecologists on the team, the class-wise mean absolute error (MAE) indicating the counting accuracy of each species:

$$\text{MAE} = \frac{1}{I}\sum_{i=1}^{I}\sum_{c=1}^{C}|\hat{n}_{i,c} - n_{i,c}|\,, \tag{3}$$

Where I is the number of images, C is the number of classes, $\hat{n}_{i,c}$ and $n_{i,c}$ are the predicted and ground truth counts for class c in image i.

Implementation Details: Each image taken from the drone is $8256 \times 5506 \times 3$. We create tiles for each image in the spatial domain, with the size of $1024 \times 1024 \times 3$ with 25% of overlap. The Patch Embed layer uses a single CNN layer with a large kernel of size 16×16 with stride 16. In the parallel branch, the *High-Frequency Feature Generator*, we use DFT to compute the Fourier transform, the mask is a binary disk with the radius set to 128. The HFC Embed layer uses 3 CNN layers with ReLU activation with a kernel of size 3×3 and a global average pool at the end. The Feature Refiner (**FR**) module consists of one cross attention layer with 1 linear layer. The image encoder is a pre-trained ViT model [21] with 24 transformer layers and 16 heads. The Query Refiner (**QR**) module takes in 100 queries each of channel dimension 256, those are cross attended with hfc_{emb} output. The box decoder contains 3 layers of two-way attention with 8 heads. We train WM with AdamW optimizer [31] setting the learning rate to 10^{-4} for the **FR**, **QR** and box decoder with a weight decay to 10^{-4}. We set the learning rate for the Patch Embed and HFC Embed layer to 10^{-5}. We load the image encoder with pre-trained weights from segment anything [21] and keep it frozen.

Data Augmentation: In the train and test datasets, we incorporated an equal number of images without any objects to assess the model's robustness against empty background images. We applied multiple data augmentation techniques, including HSV (hue, saturation, and value) (10%), rotation (5%), translation (10%), affine transformation (20%), scale (10%), shear (5%) and mosaic (70%) augmentation [1]. Mosaic augmentation is proven to be the most effective, with an improvement of 0.07 mAP

Hard Negative Mining: After training for 100 epochs, we take all the False Positives (FP) predictions having IOU≤ 0.10 with ground truth box and mark them as background class. Then fine-tuning for 20 epochs improved the performance of FP reduction for detecting rocks, dead tress or other artifacts on ground as animal.

6. Results

6.1. Performance Comparison

We trained WM separately on Mara-Wildlife dataset and Virunga-Garamba-AED dataset for comparison with existing works, see Table 2 for a summary of the results on all of the tested datasets. The *mAP* values are compared for *IoU* of 0.50-0.95 and 0.50. We also provide the average counting error in animal counting per image. We trained all the baseline models with the default set of conditions on Mara-Wildlife dataset. We merged the Virunga, Garamba, and AED datasets and created the train-val-test split according to [7]. The combination of these datasets contains 6 unique animal species and diversity of landscapes such as woodland, savannahs, open shrubland, and grasslands across multiple countries – Democratic Republic of Congo (DRC), Botswana, Namibia, and South Africa. **WM outperforms all methods by a significant margin as shown in Table 2.** We note that in [7] the authors did not make the code base or trained model public, hence we could not verify the results. We implemented these methods from

	Methods	#epochs	mAP	mAP50	Counting Error
	Mara Wildlife Dataset				
1	Faster-Rcnn [40]	100	0.24	0.58	2.59
2	DETR [3]	200	0.22	0.57	2.75
3	Co-DETR-R50 [55]	100	0.27	0.66	2.72
4	Co-DETR-swingL [55]	100	0.28	0.65	2.60
5	Yolo v5 [10]	100	0.30	0.67	2.12
6	Yolo v8 [17]	100	0.27	0.61	3.97
7	LSKNet [29]	100	0.29	0.66	-
6	DroneDetect [50]	100	0.18	0.48	-
8	**WildlifeMapper**	120	**0.56**	**0.79**	**1.9**
	Virunga-Garamba-AED Datasets				
6	Faster-Rcnn	120	0.34	0.65	0.27
7	DETR	200	0.30	0.62	0.45
8	Yolo v5	100	0.48	0.78	0.12
9	Yolo v8	100	0.48	0.77	0.42
10	WildlifeMapper	80	**0.68**	**0.85**	**0.11**

Table 2. *Comparison with baseline models. The top section shows performance on species detection on Mara-wildlife dataset and low section shows performance on the mixed dataset from Virunga-Garamba-AED dataset. The overall detection accuracy is generally higher in Virunga-Garamba-AED dataset because there are only 6 species and the terrain is quite similar in all images.*

	HFG	FR	QR	mAP
WM	✗	✗	*	0.46
	✓	✓	*	0.54
	✓	✗	✓	0.49
	✓	✓	✓	**0.56**

Table 3. *HFG module effectiveness in refining the image features and queries. "✗" represents not used, "✓" represents used and "*" represents that random queries are used but there was not refining with HFG features.*

the original public repositories and trained according to the training strategy detailed in [7]. We attribute the model poor learning performance due to salient features of the homogeneous background being learned more than the object of interest. The detection of the object of interest is then dependent on the landscape properties instead of object properties. Hence when used on a slight variations of landscapes for the same object, the models struggle to detect. This limitation is specifically addressed in the WM, where the **HFG** modules suppresses the background and highlights the object of interest.

Qualitative results: Fig. 5 shows the quality of detection by WM in different scenarios. Those include, detection when animal is partially visible under a tree, or a big clustering. WM makes correct predictions in varying scenarios.

6.2. Ablation Studies

We performed all ablation experiments on Mara-Wildlife (MW) dataset and validate the design choices.

High-frequency Feature Generator Module: In Table 3, we show the effectiveness of the **HFG** module. We experimented with **HFG**'s output in 3 ways: first, we passed **HFG**'s output to Feature Refiner (**FG**) module. It leads to significant improvement in detection by 0.09 mAP over the baseline. This demonstrate that providing potential location candidates features to image encoder module produce better embeddings. Second, we pass the **HFG**'s output to Query Refiner (QR) module only. This leads to an improvement of 0.03 mAP over baseline. This shows the effectiveness of guiding queries with location candidates features. In the third case, we passed the **HFG**'s output to both **FR** and **QR** modules and achieved an improvement of 0.11 mAP over the baseline. We hypothesize that this reduces the dominance of features from homogeneous and dominant background in aerial imagery. The also observed this while testing WM across flightlines different types of terrain such as green grasslands, dry grasslands, and forest areas.

Feature Refiner Module: We tested hfc_{emb} and f_{img} merging by 3 ways: addition, concatenation and cross-attention. Cross-attention is most effective, because with addition and concatenation, the hfc_{emb} get lost, while cross-attention generates better embeddings giving attention to potential location candidates.

Kernel Size: We observed that a larger kernel size of 31×31 results in reduced misclassification. For example, a $topi$ or $warthog$ cannot be found inside a $boma$ because only domestic species are kept in $bomas$. So context helps in making the right class detection. We observed an improvement of 0.02 mAP. Experiments were done in 3 kernel sizes. **1.** 7×7: 0.55 mAP, **2.** 16×16: 0.558 mAP; **3.** 32×32: 0.57 mAP.

Query Refiner Module: We experimented with only providing random queries and guiding the queries with **HFG** module output. We merged them with direct addition or concatenation and cross-attention. With cross-attention, we observed an improvement in performance of 0.03mAP.

Geographic generalization: To test the geographic generalizability of WM across different terrains, we trained WM only on images from Kenya; and tested on images from Democratic Republic of Congo, Botswana and Namibia. The test was done on 4 common species which were present in both the ecosystems. WM achieved a detection performance of 0.48 mAP. This shows the adaptability of WM across varying landscapes.

Domain generalization: We train-test WM on a different domain, a bird species tern [47], commonly found on/near water bodies. Live in huge clusters. WM achieved the accuracy of 0.71 mAP. Showing adaptability of WM across different domains.

Figure 5. *Qualitative results. The top row highlights examples of crowded and partially occluded scenes. Row-1, Column-1 & Row-3, Column-3 shows examples where animals are obstructed by shadows. The zoomed-in box in Row-3, Column-3 shows a zebra partially occluded. The bounding box color is coded according to class names: shoats-"hot pink", cattle-"deep sky blue", zebra-"light yellow".*

Figure 6. *Failure cases. Left shows an example where animals are occluded by shadows and are difficult to detect. Right shows an example of rock detected as an impala, emphasizing the difficulty in differentiating objects in the image from animals of interest.*

Failure Cases: The detections from WM are inaccurate when animals are clustered in shadows, such as when animals are located inside bomas and the sun angle makes a strong shadow on the enclosure. These are some of the difficult cases shown in Fig. 6. Other cases of false positives are the small rocks or trees that sometimes resemble animals. Some examples are shown in the supplementary materials.

7. Conclusion

This paper presents WildlifeMapper (WM) - a transformer based approach for the detection of animals of varying densities and sizes across natural backgrounds. The WM utilizes a high frequency features generator, feature refiner, and query refiner to accurately locate and classify 8 animal species. WM stands to significantly improve the efficiency and accuracy of wildlife monitoring and conservation efforts. Future work will extend this model and dataset to a larger number of species and habitats.

8. Acknowledgement

This research is partially supported by the following grants: NSF award SI2-SSI #1664172 (BSM), the Ohrstrom Family Foundation, and a Smithsonian National Zoo and Conservation Biology Institute (SNZCBI) fellowship (S.Kumar). We also recognize the contributions of the annotation team, including S. Harman, A. Gray, C. Pate, C. Obath, S. Nyarangi, C. Kagume, and W. Sairowua.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 3, 6

[2] Fergus W Campbell and John G Robson. Application of fourier analysis to the visibility of gratings. *The Journal of physiology*, 197(3):551, 1968. 5

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3, 5, 6, 7

[4] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021. 3

[5] Qiang Chen, Jian Wang, Chuchu Han, Shan Zhang, Zexian Li, Xiaokang Chen, Jiahui Chen, Xiaodi Wang, Shuming Han, Gang Zhang, et al. Group detr v2: Strong object detector with encoder-decoder pretraining. *arXiv preprint arXiv:2211.03594*, 2022. 3

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 5

[7] Alexandre Delplanque, Samuel Foucher, Philippe Lejeune, Julie Linchant, and Jérôme Théau. Multispecies detection and identification of african mammals in aerial imagery using convolutional neural networks. *Remote Sensing in Ecology and Conservation*, 8(2):166–179, 2022. 1, 2, 3, 6, 7

[8] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7778–7796, 2021. 2

[9] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022. 4

[10] Glenn Jocher et. al. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, Oct. 2020. 7

[11] Emilio Guirado, Siham Tabik, Marga L Rivas, Domingo Alcaraz-Segura, and Francisco Herrera. Whale counting in satellite and aerial images with deep learning. *Scientific reports*, 9(1):14259, 2019. 1

[12] Liang Han, Pin Tao, and Ralph R Martin. Livestock detection in aerial images using a fully convolutional network. *Computational Visual Media*, 5:221–228, 2019. 1, 3

[13] Tracey Hollings, Mark Burgman, Mary van Andel, Marius Gilbert, Timothy Robinson, and Andrew Robinson. How do you find the green sheep? a critical review of the use of remotely sensed imagery to detect and count animals. *Methods in Ecology and Evolution*, 9(4):881–892, 2018. 2

[14] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal networks. In *The IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017. 2

[15] ASM Iftekhar, Satish Kumar, R Austin McEver, Suya You, and BS Manjunath. Gtnet: Guided transformer network for detecting human-object interactions. In *Pattern Recognition and Tracking XXXIV*, volume 12527, pages 192–205. SPIE, 2023. 5

[16] Johannes Jakubik et. al. Prithvi-100M, Aug. 2023. 3

[17] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, Jan. 2023. 7

[18] Benjamin Kellenberger, Diego Marcos, and Devis Tuia. Detecting mammals in uav images: Best practices to address a substantially imbalanced dataset with deep learning. *Remote Sensing of Environment*, 216:139–153, Oct 2018. 1, 2

[19] Benjamin Kellenberger, Devis Tuia, and Dan Morris. Aide: Accelerating image-based ecological surveys with interactive machine learning. *Methods in Ecology and Evolution*, 11(12):1716–1727, 2020. 4

[20] Benjamin Kellenberger, Michele Volpi, and Devis Tuia. Fast animal detection in uav images using convolutional neural networks. In *2017 IEEE international geoscience and remote sensing symposium (IGARSS)*, pages 866–869. IEEE, 2017. 2

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 4, 5, 6

[22] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[23] Satish Kumar, Ivan Arevalo, ASM Iftekhar, and BS Manjunath. Methanemapper: Spectral absorption aware hyperspectral transformer for methane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17609–17618, 2023. 3, 4, 5

[24] Satish Kumar, William Kingwill, Rozanne Mouton, Wojciech Adamczyk, Robert Huppertz, and Evan D Sherwin. Guided transformer network for detecting methane emissions in sentinel-2 satellite imagery. In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*, 2022. 3

[25] Kristian Kvilekval, Dmitry Fedorov, Boguslaw Obara, Ambuj Singh, and BS Manjunath. Bisque: a platform for bioimage analysis and management, Jan. 2023. 2, 4

[26] Richard Lamprey, David Ochanda, Rob Brett, Charles Tumwesigye, and Iain Douglas-Hamilton. Cameras replace human observers in multi-species aerial counts in murchison falls, uganda. *Remote Sensing in Ecology and Conservation*, 6(4):529–545, 2020. 2

[27] Richard Lamprey, Frank Pope, Shadrack Ngene, Michael Norton-Griffiths, Howard Frederick, Benson Okita-Ouma, and Iain Douglas-Hamilton. Comparing an automated high-definition oblique camera system to rear-seat-observers in a wildlife survey in tsavo, kenya: taking multi-species aerial counts to the next level. *Biological Conservation*, 241:108243, 2020. 1

[28] Richard Lamprey, Frank Pope, Shadrack Ngene, Michael Norton-Griffiths, Howard Frederick, Benson Okita-Ouma,

and Iain Douglas-Hamilton. Comparing an automated high-definition oblique camera system to rear-seat-observers in a wildlife survey in tsavo, kenya: Taking multi-species aerial counts to the next level. *Biological Conservation*, 241:108243, 2020. 2

[29] Yuxuan Li, Qibin Hou, Zhaohui Zheng, Ming-Ming Cheng, Jian Yang, and Xiang Li. Large selective kernel network for remote sensing object detection. *arXiv preprint arXiv:2303.09030*, 2023. 3, 7

[30] Julie Linchant, Jonathan Lisein, Jean Semeki, Philippe Lejeune, and Cédric Vermeulen. Are unmanned aircraft systems (uas s) the future of wildlife monitoring? a review of accomplishments and challenges. *Mammal Review*, 45(4):239–252, 2015. 2

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[32] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021. 3

[33] Johannes Naude and Deon Joubert. The aerial elephant dataset: A new public benchmark for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 48–55, 2019. 3

[34] Joseph O Ogutu, Hans-Peter Piepho, Mohamed Y Said, Gordon O Ojwang, Lucy W Njino, Shem C Kifugo, and Patrick W Wargute. Extreme wildlife declines and concurrent increase in livestock numbers in kenya: What are the causes? *PloS one*, 11(9):e0163249, 2016. 1

[35] Wilber K Ottichilo, Jesse Grunblatt, Mohammed Y Said, and Patrick W Wargute. Wildlife and livestock population trends in the kenya rangeland. *Wildlife conservation by sustainable use*, pages 203–218, 2000. 1

[36] Wilber K. Ottichilo, Jesse Grunblatt, Mohammed Y. Said, and Patrick W. Wargute. *Wildlife and Livestock Population Trends in the Kenya Rangeland*, pages 203–218. Springer Netherlands, Dordrecht, 2000. 2

[37] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 5

[38] Michael John Stephen Peel and Marc Stalmans. The systematic reconnaissance flight (srf) as a tool in assessing the ecological impact of a rural development programme in an extensive area of the lowveld of south africa. *African Journal of Ecology*, 37(4):449–456, 1999. 1

[39] Tinao Petso, Rodrigo S Jamisola Jr, Dimane Mpoeleng, Emily Bennitt, and Wazha Mmereki. Automatic animal identification from drone camera based on point pattern analysis of herd behaviour. *Ecological Informatics*, 66:101485, 2021. 1, 2

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3, 7

[41] Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv preprint arXiv:2111.14330*, 2021. 3

[42] AC Seymour, J Dale, M Hammill, PN Halpin, and DW Johnston. Automated detection and enumeration of marine wildlife using unmanned aircraft systems (uas) and thermal imagery. *Scientific reports*, 7(1):45127, 2017. 1

[43] Wim Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM journal on mathematical analysis*, 29(2):511–546, 1998. 5

[44] Colin J Torney, David J Lloyd-Jones, Mark Chevallier, David C Moyer, Honori T Maliti, Machoke Mwita, Edward M Kohi, and Grant C Hopcraft. A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods in Ecology and Evolution*, 10(6):779–787, 2019. 1, 2

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 5

[46] Cédric Vermeulen, Philippe Lejeune, Jonathan Lisein, Prosper Sawadogo, and Philippe Bouché. Unmanned aerial survey of elephants. *PloS one*, 8(2):e54700, 2013. 2

[47] Ben G Weinstein, Lindsey Garner, Vienna R Saccomanno, Ashley Steinkraus, Andrew Ortega, Kristen Brush, Glenda Yenni, Ann E McKellar, Rowan Converse, Christopher D Lippitt, et al. A general deep learning model for bird detection in high-resolution airborne imagery. *Ecological Applications*, 32(8):e2694, 2022. 7

[48] Zijing Wu, Ce Zhang, Xiaowei Gu, Isla Duporge, Lacey F Hughey, Jared A Stabach, Andrew K Skidmore, J Grant C Hopcraft, Stephen J Lee, Peter M Atkinson, et al. Deep learning enables satellite-based monitoring of large populations of terrestrial mammals across heterogeneous landscape. *Nature communications*, 14(1):3072, 2023. 2, 3

[49] Xiangkai Xu, Zhejun Feng, Changqing Cao, Mengyuan Li, Jin Wu, Zengyan Wu, Yajie Shang, and Shubing Ye. An improved swin transformer-based model for remote sensing object detection and instance segmentation. *Remote Sensing*, 13(23):4779, 2021. 3

[50] Yinhui Yu, Xu Sun, and Qing Cheng. Expert teacher based on foundation image segmentation model for object detection in aerial images. *Scientific Reports*, 13(1):21964, 2023. 7

[51] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021. 5

[52] Jiahe Zhu, Xu Chen, Huan Zhang, Zelong Tan, Shengjin Wang, and Hongbing Ma. Transformer based remote sensing object detection with enhanced multispectral feature extraction. *IEEE Geoscience and Remote Sensing Letters*, 2023. 3

[53] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018. 2

[54] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3

[55] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023. 7