Supplementary Material
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

1. Introduction

We provide all the supplementary information related to WildlifeMapper (WM) and the Mara-Wildlife (MW) dataset here. We also provide qualitative examples.

2. Method

High Frequency Feature Generator (HFG): This section covers the detailed derivation and implementation information of our **HFG** module. The input image is processed in parallel by the **HFG** module to generate features with information about the location of the animal or cluster. The **HFG** module is inspired from the limitation of ViT models [6]. 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 [4, 7].

Research indicates that local features in images are closely linked to high-frequency components [1, 5]. 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..

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 ($IDFT$) to get back image I'. The DFT is computed as:

$$F(u,v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} I(x,y) \cdot e^{-j2\pi(ux/H + vy/W)} \tag{1}$$

where $F(u,v)$ is the magnitude spectrum, u and v are the frequency coordinates, and j is the imaginary unit. Next, we shift the lower frequency components to center of frequency spectrum as:

$$F'(u,v) = F\left((u + \frac{H}{2}) \mod H, (v + \frac{W}{2}) \mod W\right), \tag{2}$$

where mod is modulus operation. Next we mask the lower frequencies with a controlling parameter r. The modified Fourier transform G with mask M is defined as:

$$G(u,v) = M(u,v) \odot F'(u,v) \tag{3}$$

$$\text{where } M = \begin{cases} 1 & \text{if} (u - H/2)^2 + (v - W/2)^2 \leq r^2 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Given the modified Fourier transform G, the reconstructed image I' is given by:

$$I'(x,y) = \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} G(u,v) \cdot e^{j2\pi(ux/H + vy/W)} \tag{5}$$

Next we reduce the dimension of the reconstructed image $I'(x,y)$ via an embedding layer to generate embedding hfc_{emb} and pass them to the **FR** module

2.1. 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 [2] 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 [3] 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 [2] and keep it frozen.

3. Mara-Wildlife Dataset:

Flight path details The chosen flight path prioritized vast open grasslands, as they frequently serve as habitats and transit routes for larger fauna. The survey was conducted in March. March is typically a rainy month when grasses are green. The Serengeti migratory herd of wildebeest have already moved south from the region. Data collection was typically scheduled during the early mornings or late afternoons. These times are when animals are most active, avoiding the midday sun. Although the evening presents challenges due to diminished sunlight, the majority of the data was acquired between 7AM and 10AM local time to ensure optimal lighting conditions.

Camera Settings and Specifications: We mounted a NIKON D850 camera to the bellyport of the airplane. The camera was placed in a NADIR view and configured with an intervalometer to collect an image every two seconds along flight transects.

3.1. Rich Metadata for Computer Vision Benchmarks:

Each raster included in the dataset is accompanied with detailed metadata, the timestamp of the image capture, and other camera EXIF (Exchangeable Image File Format) information such as focal length, FNumber, ISO, and ExposureTime. We collected latitude and longitude and elevation information from the GPS log of the pilot and merged this information using the camera data and time. If released with the data, these metadata properties enrich the dataset's ecological value and unlock the potential for a myriad of computational applications.

While the primary intent of our analysis was to provide an estimate of the abundance of large mammals across the Masai Mara ecosystem, the dataset's comprehensive nature presents opportunities that extend beyond wildlife studies. These include: *Sun Angle Prediction*, *Image Registration*, *GPS Estimation*, *Elevation Prediction*.

- *Sun Angle Prediction:* Given the timestamp and known location of each image capture, the dataset could be employed to develop models that predict the sun's angle based on the image content. Such applications can benefit fields ranging from photovoltaic systems to architectural planning.
- *Image Registration:* The dataset provides a platform for researchers to work on algorithms that align or 'register' multiple images of the same region, even if taken from varying angles or times. Such tasks find relevance in areas like medical imaging and satellite image analysis.
- *GPS Estimation:* The precise latitudinal and longitudinal coordinates embedded in the metadata allow for the creation of models that predict the GPS location of specific objects or even individual pixels using only the image content. This potential extends the bounds of localization models in the realm of computer vision.
- *Elevation Prediction:* The dataset's rich elevation data provides an avenue to train models that can estimate the altitude at which an image was taken, based purely on visual cues. Such models can have vast applications, from aviation to drone technology.

These represent just a few of the many potential applications. We believe the Mara-Wildlife dataset has the potential to be a foundational resource for both ecological studies and computer vision research, ushering in innovations and novel solutions.

4. Results

In this section, we present more qualitative results of detection of WildlifeMapper on the Mara-Wildlife dataset. Good detection samples are shown in Fig. 1 and the failure cases are shown in Fig.

Good cases: Each column in Fig. 1 shows detections of different types of animals. Column-1 shows large animals: $cattle, buffalo$; Column-2 shows detection of small animals: $warthog, topi$. Column-3 shows detection of animals hidden or occluded in Row 1 & 2, Row 3 & 4 show examples of $other$ categories (i.e., lion).

Failure cases: Fig. 2 shows examples of where WildlifeMapper struggled to make an accurate detection. Each image shows a unique scenario where the detection was either missed or misclassified or confused from the contextual information. For example in Column-1, Row-1, the dry wooden log is detected as an object and misclassified as $shoat (sheep\ or\ goat)$ since $shoats$ almost always occur in a group. Hence, the additional object identified was mislabeled a $shoat$. ***Figures are on next page.***

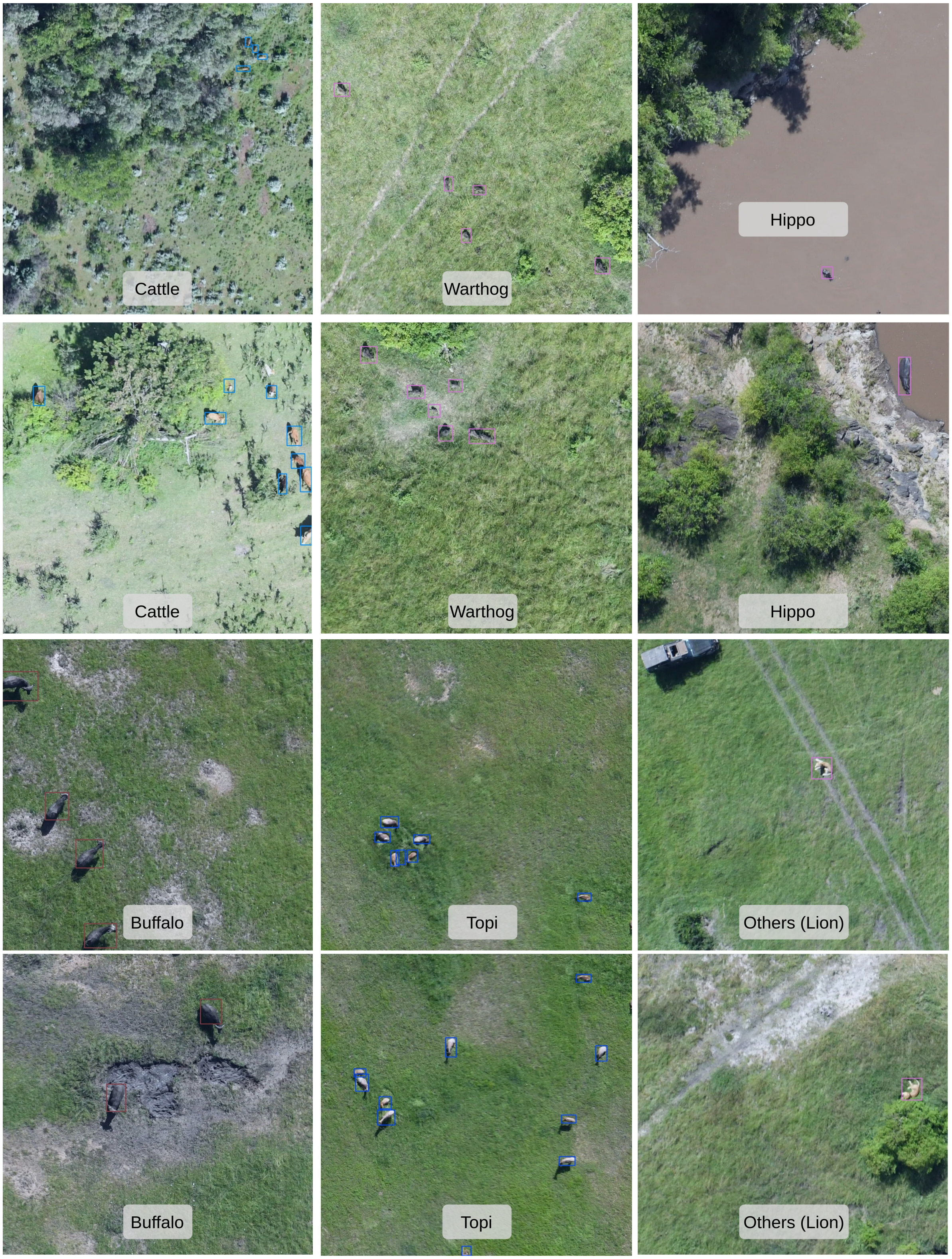

Figure 1. *Good cases. Each column shows different category of detection. Column-1 shows large animals: cattle, buffalo; Column-2 shows detection of small animals (warthog, topi), Column-3 shows detection of animals hidden or occluded.*

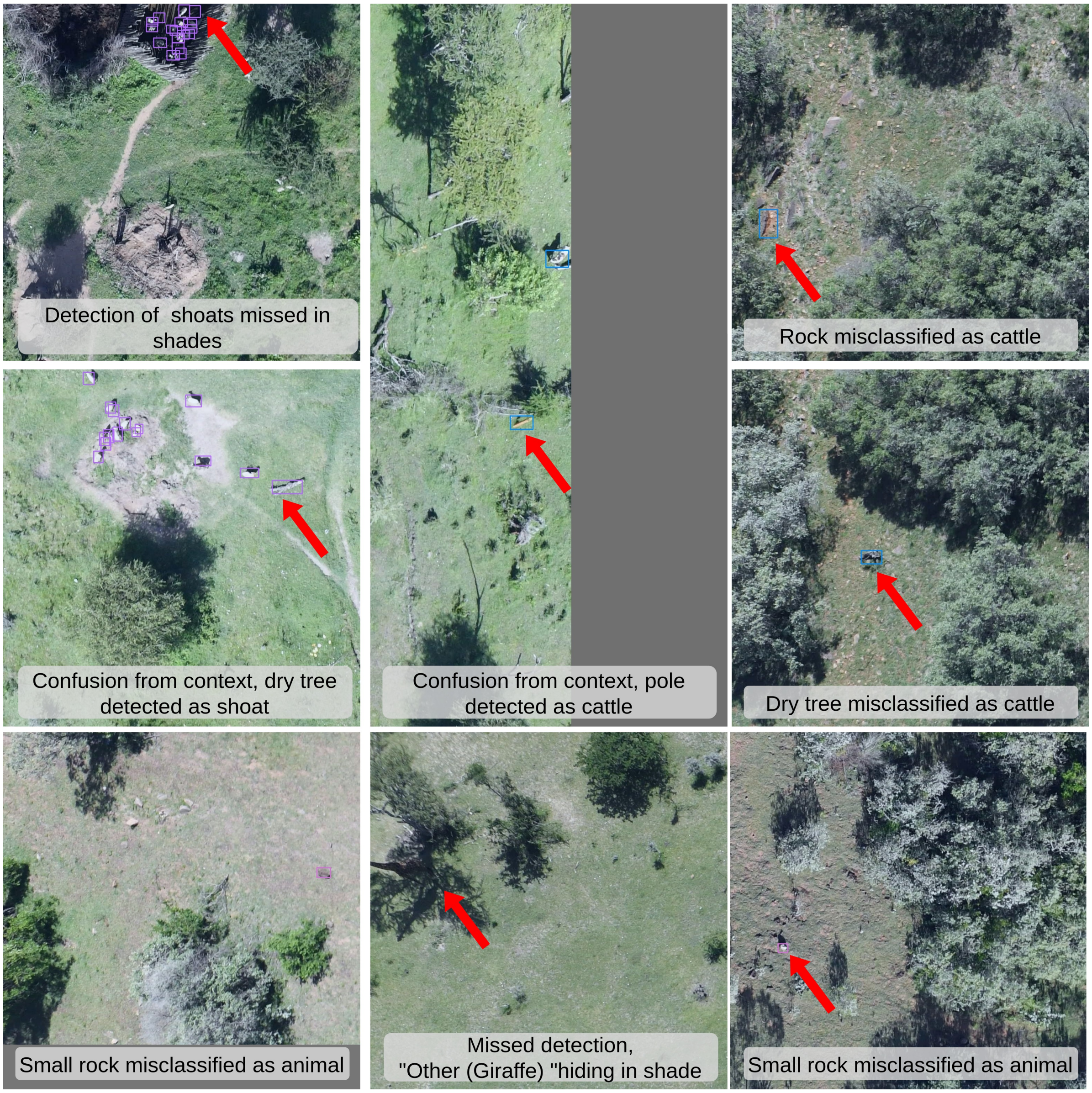
Detection of shoats missed in shades
Confusion from context, pole detected as cattle
Rock misclassified as cattle
Confusion from context, dry tree detected as shoat
Dry tree misclassified as cattle
Small rock misclassified as animal
Missed detection, "Other (Giraffe) "hiding in shade
Small rock misclassified as animal

Figure 2. *Failure cases. The animals hiding in the shade are difficult to detect. Additional examples of misclassification also provided.*

References

[1] Fergus W Campbell and John G Robson. Application of fourier analysis to the visibility of gratings. *The Journal of physiology*, 197(3):551, 1968. 1

[2] 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. 2

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[4] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 1

[5] Wim Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM journal on mathematical analysis*, 29(2):511–546, 1998. 1

[6] 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. 1

[7] 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. 1