InfraNet: An Ensemble Approach for Real-time Wildlife Detection using Infrared Thermal Imaging

Dheeraj Dhillon, Vinod Pankajakshan
Department of Electronics & Communication Engineering
Indian Institute of Technology Roorkee
dheeraj_d@ec.iitr.ac.in, vinod.pankajakshan@ece.iitr.ac.in

Parvathi M S, Sreejith Sajeev, Joby Thomas, Byju C, Rajesh K R
Strategic Electronics Group
C-DAC Thiruvananthapuram, India
parvathi.ms@cdac.in, segosengr10@cdac.in, joby@cdac.in, byjuc@cdac.in,rajesh@cdac.in

Abstract

Human-wildlife conflict presents significant challenges to both conservation and human safety, necessitating efficient monitoring systems for timely wildlife detection. We introduce InfraNet, an infrared object detection system designed for real-time wildlife monitoring using embedded devices. Our key contributions are (1) a new annotated infrared dataset of elephants, human, and common domestic animals, curated to capture diverse environmental conditions, and (2) an ensemble methodology that combines predictions of multiple preprocessed thermal image versions using a baseline YOLOv8m model without fine-tuning. Experimental results on a set of Elephant dataset show that the proposed ensemble approach significantly increases recall rate from 0.35 to 0.62. Additionally, the ensemble model achieves real-time inference speeds on an NVIDIA Jetson Xavier NX, making it suitable for field deployment.

1. Introduction

Human-wildlife conflict has become a growing concern globally, posing significant challenges to both wildlife conservation and human communities. As human populations expand and encroach upon natural habitats, interactions between humans and wildlife have increased, often with disastrous outcomes. For instance, in Africa and Asia, elephants frequently raid crops, causing economic losses for farmers and sometimes resulting in human fatalities. In retaliation, communities may harm or kill wildlife, further threatening already endangered species. In India, the situation is severe. The country is home to over 50% of Asia's wild elephants [12], and incidents of human-elephant conflict are increasing. Villagers have taken extreme measures such as electric fencing and poisoning to protect their livelihoods, emphasizing the need for effective mitigation strategies. Traditional methods to prevent human-wildlife conflicts, such as physical barriers, relocation of animals, or community-based deterrents, have had limited success due to animal adaptability and habitat fragmentation [13]. Innovative solutions are needed to proactively prevent encounters. Early detection of wildlife approaching human settlements is crucial for implementing timely interventions. Advanced monitoring systems using seismic [14], [2] and imagery data can significantly reduce the risks associated with human-wildlife conflicts, but designing such systems poses challenges in remote or resource-constrained environments. Infrared thermal imaging has emerged as a valuable tool in wildlife monitoring due to its ability to capture thermal signatures of animals regardless of lighting conditions or other environmental factors [10]. Thermal imaging enables continuous surveillance by detecting animals even at night, unlike visible-light cameras that rely on adequate illumination. Thermal cameras have been successfully used to monitor nocturnal movements of elephants in African savannas [4], leading to a better understanding and management of their interactions with human settlements. Despite its potential, infrared object detection for wildlife monitoring is fraught with several challenges. In tropical regions, high ambient temperatures reduce thermal contrast, making it challenging to distinguish animals from their surroundings in infrared imagery. Models trained on visible images often fail to generalize to infrared imagery because of fundamental differences in texture, contrast, and feature representation. Furthermore, infrared sensors may exhibit decreased sensitivity and increased noise under extreme conditions, affect-

ing image quality and thereby the detection reliability. Environmental factors like humidity, foliage density, and atmospheric conditions also introduce noise and artifacts in the thermal images. Thermal signatures vary by time of day, generally improving at night when temperatures are cooler but require systems to perform reliably under varying conditions. Researchers have also applied deep learning techniques to infrared imagery, often retraining models on infrared datasets to improve performance. For instance, Zhou et al. [16] introduced YOLO-CIR, which combines YOLO with ConvNeXt, achieving better results on standard thermal image datasets. Retraining these models on infrared datasets is resource-intensive and often impractical due to the scarcity of annotated infrared data and limited computational resources. Furthermore, the lack of large annotated infrared datasets makes it challenging to train or retrain deep learning models specifically for infrared imagery. In this work, we introduce *InfraNet*, an infrared thermal image object detection system designed for real-time wildlife monitoring on embedded devices. Our motivation arises from the need to develop a functional and efficient system for early wildlife detection to mitigate human-wildlife conflicts. By enhancing detection accuracy in infrared images without fine-tuning the trained detection models, we aim to provide a solution that can be readily deployed in environments with limited computational resources, ultimately benefiting both wildlife conservation and human safety. The proposed approach leverages preprocessing techniques and an ensemble methodology to enhance detection accuracy using a single pretrained model. By focusing on infrared-only detection and optimizing for efficiency, InfraNet provides a practical solution for early wildlife detection in areas affected by human-wildlife conflict, allowing for timely interventions that promote both wildlife conservation and human safety. This work focuses specifically on designing an elephant monitoring system using thermal infrared imaging. The paper is structured as follows: Section 2 presents a new annotated thermal image dataset, Section 3 details the proposed methodology, Section 4 discusses the experimental results, and Section 5 concludes the paper.

2. Dataset

The advancement of object detection in thermal images relies significantly on the availability of annotated thermal datasets. Public datasets include the FLIR ADAS Thermal Dataset [3], the KAIST Multispectral Pedestrian Detection Benchmark [5], and the LLVIP dataset [6]. Although these datasets are helpful for urban object detection involving pedestrians, vehicles, and other related objects, their applicability to wildlife monitoring is limited. We introduce a new annotated infrared thermal image dataset collected from regions experiencing human-elephant conflicts in India [1]. The dataset provides a comprehensive collection of annotated infrared images of elephants captured in diverse environmental settings, including varying temperatures, backgrounds, and times of day. This variation makes it suitable for evaluating the generalization capability of various infrared object detection models. Data were collected from 3 different locations: Chilla forest range of Uttarakhand, the Elephant Training Center in Kerala, and Trivandrum, Kerala. The datasets include $50,694$ images and $112,816$ annotated instances across three classes: *elephant* (class_0), *person* (class_1), and *other* (class_2), encompassing common domestic animals goat, dog, cow, and horse. The images were extracted from thermal videos with spatial resolution 640×512 pixels at 10 frames per second. A FLIR Boson thermal camera having pixel size $12\ \mu$m, 24° field of view, and focal length of 18 mm was used to record the videos. Table 1 summarizes the attributes of the thermal datasets. The Chilla dataset, collected during 11:00 am to 6:00 pm, with temperature varying from 38 to 45°C, has challenging thermal contrast. The Konni dataset, recordeded during 8:00 am to 10:00 am, presented cooler temperatures (25–30°C) and varied terrain, providing differing environmental conditions. The Trivandrum dataset, collected during 7:00 am to 6:00 pm, predominantly consists of domestic animal images, provides negative samples to evaluate false-positive detection performance. For each image in the dataset, there is a corresponding label file containing bounding box annotations in the YOLO label format. The class distribution is imbalanced, with an abundance of *person* instances in the Konni dataset and numerous *other* animal instances in the Trivandrum dataset. Figure 1 illustrates the dataset's diversity through sample labeled thermal images, showcasing various backgrounds, animal poses, and environmental conditions.

3. Methodology

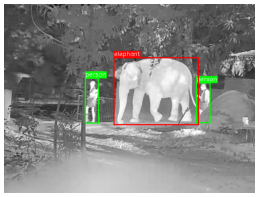
We employ a deep learning architecture with real-time object detection capability, essential for timely wildlife monitoring. This section outlines the proposed methodology, including baseline model setup, preprocessing techniques, ensemble approach, and the prediction aggregation process.

3.1. Baseline Model

In this work, we choose the popular object detection algorithm YOLOv8m [7] as the baseline model. This model is pre-trained on the COCO-2017 dataset [8] which consists of normal visible light images. The model is trained to classify 80 classes, including elephant and person. To adapt the pretrained model to our custom use case with 3 classes (elephant, person, and other), we map domestic animal classes (horses, goats, cows, and dogs) to the "other" category and ignore remaining class predictions. This baseline serves as

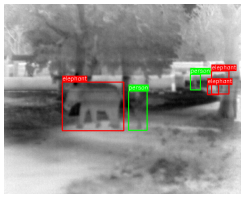
Attribute	Chilla	Konni	Trivandrum	Total
Total Images	6,289	37,356	7,049	50,694
Instances Images	6,289	28,214	6,603	41,106
Background Images	0	9,142	446	9,588
Total Instances	12,175	92,814	7,827	112,816
Elephant Images	6,101	16,337	0	22,438
Elephant Instances	11,306	20,604	0	31,910
Person Instances	864	72,210	866	73,940
Other Instances	5	0	6,961	6,966
Temp Range (°C)	38–45	25–30	25–30	–
Capture Timeframe	11 am - 6 pm	8 am - 10 am	7 am - 6 pm	–

Table 1: Attributes of the thermal dataset

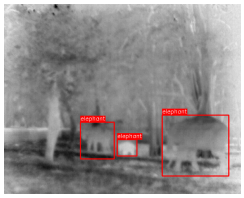


(a) Konni, Kerala

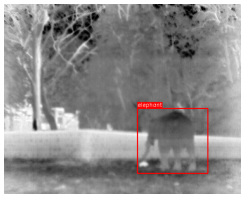
(b) Konni, Kerala



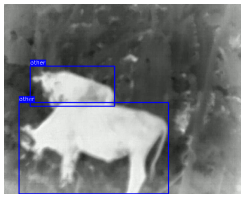
(c) Chilla, Uttarakhand



(d) Chilla, Uttarakhand



(e) Chilla, Uttarakhand



(f) Trivandrum, Kerala

Figure 1: Sample thermal images with labels

a benchmark for evaluating different approaches on our collected dataset, particularly assessing its ability to generalize on infrared thermal image dataset without fine-tuning.

3.2. Thermal Image Preprocessing

Our initial experiments with fine-tuning the baseline model using thermal images indicated that the model overfits the training set and performs poorly on the validation set, particularly in the challenging Chilla dataset. Hence we explore various pre-processing steps to improve the performance of the baseline model instead of fine-tuning the baseline model. To improve generalization, we preprocess the raw thermal frames, which are in grayscale format, before feeding them into the detection model. We consider the following two preprocessing techniques.

3.2.1 Image Inversion

Image inversion transforms each pixel intensity $I(x, y)$ to its negative $I_{\text{inv}}(x, y)$:

$$I_{\text{inv}}(x, y) = 255 - I(x, y). \tag{1}$$

The rationale behind this preprocessing step is that it improves the contrast of white or gray details in dark backgrounds, as in the case of the Chilla dataset, where the elephants appear darker than the background due to higher ambient temperature.

3.2.2 Bilateral Filtering

Bilateral filtering is an edge-preserving and noise-reducing smoothing technique [9, 15]. This filtering reduces noise while preserving important edges, which is crucial for detecting faint edges in infrared images. It computes a weighted average of nearby pixels based on spatial proximity and intensity similarity:

$$I_{\text{bf}}(x, y) = \frac{1}{W_p} \sum_{(i,j)\in S} G_s\left(\|(i, j) - (x, y)\|\right) \times G_r\left(|I(i, j) - I(x, y)|\right) \cdot I(i, j), \tag{2}$$

where G_s is the spatial Gaussian kernel, G_r is the range Gaussian kernel, W_p is the normalization factor, and S is the spatial neighborhood. The bilateral filtering parameters used in our implementation are: the spatial neighborhood diameter of 9 pixels, and both the range G_r and the spatial G_s Gaussian kernels having standard deviation of 75.

3.3. Ensemble Method

To enhance detection accuracy and generalize across varying thermal signatures, we implement an ensemble approach that combines predictions from two preprocessing techniques: inversion and bilateral filtering. For each input image, we consider four variants: the original, inverted, bilaterally filtered, and bilaterally filtered version of the inverted image. These representations provide the YOLO detection model with diverse features that are more prominent under specific preprocessing conditions, thereby improving overall detection performance. Each preprocessed image is passed to the baseline YOLO model, generating distinct prediction sets and only the predictions which are having a confidence score greater than a certain threshold α are retained. In the aggregation step, a Non-Max Suppression (NMS) strategy [11] is applied to the four image prediction sets using an *Intersection over Union* (IoU) threshold θ to eliminate redundant detections. Initially, a class-wise NMS step is performed to retain only the highest confidence instance of overlapping detection of each class. Furthermore, if an *elephant* and an *other* detection overlap with an *IoU* greater than θ, the *elephant* detection is retained irrespective of comparison with the *other* instance. However, we also retain the *other* instance if it has a higher confidence score. Additionally, each detection retains information about its preprocessing method, allowing us to analyze the contribution of each technique to the final detection performance. The detailed steps of this ensemble process are outlined in Algorithm 1.

Algorithm 1 InfraNet Ensemble Detection

Require: Image I, Preprocessing techniques $\{\text{inv}, \text{bf}\}$, Detection model M, IoU threshold θ
Ensure: Final detections F

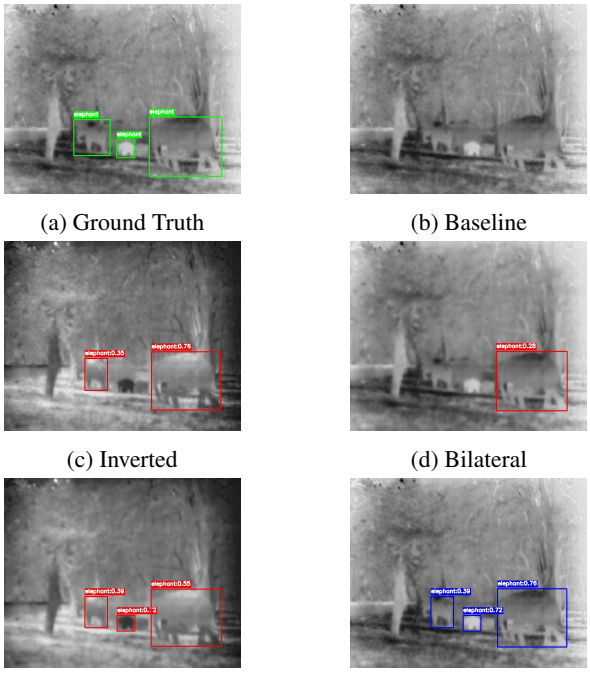
$D \leftarrow \emptyset$, $F \leftarrow \emptyset$
for each I_k in $\{I, \text{inv}(I), \text{bf}(I), \text{bf(inv(I))}\}$ **do**
  $P_k \leftarrow M(I_k)$
  $D \leftarrow D \cup \text{MapClassesCustom}(P_k)$
end for
for each class $c \in \{\text{elephant}, \text{person}, \text{other}\}$ **do**
  $D_c \leftarrow \{d \in D \mid d.\text{label} = c\}$
  Sort D_c by d.score descending
  while $D_c \neq \emptyset$ **do**
    $d \leftarrow \text{argmax}(D_c.\text{score})$
    $F \leftarrow F \cup \{d\}$
    $D_c \leftarrow D_c \setminus \{d' \mid \text{IoU}(d, d') > \theta\}$
  end while
end for
$E \leftarrow \{d \in F \mid d.\text{label} = \text{elephant}\}$
$P \leftarrow \{d \in F \mid d.\text{label} = \text{person}\}$
$O \leftarrow \{d \in F \mid d.\text{label} = \text{other}\}$
for each $elephant \in E$ **do**
  $O \leftarrow O \setminus \{o.other \in O \mid \text{IoU}(elephant, other) > \theta \wedge elephant.\text{score} \geq other.\text{score}\}$
end for
$F \leftarrow E \cup P \cup O$
return F

4. Experimental results

In this Section, we present a detailed performance analysis of the proposed ensemble method in comparison to the baseline model. The evaluation metrics considered are *precision* and *recall*, which are critical to assess object detection performance in wildlife monitoring applications. Additionally, the contributions of each preprocessing variant towards the overall recall of the ensemble model is analyzed through a detailed breakdown of True Positives (TP) and False Positives (FP). We set a minimum confidence threshold α as 0.25 for detections to filter out low-confidence predictions, ensuring a balance between true and false positives. In the aggregation step, we have used the *IoU* threshold $\theta = 0.5$ for non-max suppression.

Table 2 reports the frame-level precision and recall for each of the considered classes (*elephant*, *person*, and *other*) across different dataset locations (*Trivandrum*, *Konni*, and *Chilla*) and the overall performance. In *Trivandrum* dataset, the ensemble approach achieves a recall of 0.930 for the *other* class, increasing from 0.534 in the baseline model. The low precision values for the *person* class in *Trivandrum* dataset is attributed to the misclassifications of many *other* objects as *person* due to close proximity with the camera. Furthermore, the number of *person* instances in the dataset is very few compared to the number of misclassifications, and hence resulting in the low precision score. In the *Konni* dataset, the ensemble method increases the recall for *elephant* detections from 0.438 in baseline model to 0.573. This relatively lower recall score is due to the strict labeling carried out during the annotation process. There are many instances where the elephant is partially occluded by dense foliage, but still labeled as *elephant*. Figure 3 shows one such example where the elephant is hardly visible but still labeled as *elephant*. The most noticeable improvement in the recall rate for the *elephant* class is observed in the *Chilla* dataset, where the recall significantly increased from 0.124 in the baseline model to 0.767 in the ensemble model. As illustrated in Fig.2, the low recall rate of the baseline model is because of the low image contrast due to high ambient temperature. The baseline model fails to detect any of the elephants present in the frame, whereas the ensemble model accurately detects all three elephants.

In summary, the ensemble model consistently outperforms individual preprocessing techniques in terms of recall across all classes and locations, indicating a higher detection rate of true objects. It is to be noted that the en-



(a) Ground Truth (b) Baseline

(c) Inverted (d) Bilateral

(e) Inverted Bilateral (f) Ensemble

Figure 2: Prediction results for an image in the *Chilla* dataset

Figure 3: Example of a missed *elephant* detection.

semble model enhances the recall performance with only a marginal decrease in the precision rate. High precision implies lower false alarm rate, which is equally important as the recall for a reliable wildlife detection system. It is interesting to analyze the contributions of the predictions from the input image variants to the predictions of the ensemble model. The number of true positive detections given in Table 3 for each preprocessed variant of input image give the information about the individual contributions to the overall recall of the ensemble model outputs. For instance, in the Chilla dataset, the major contribution to recall is coming from the inverted and bilaterally filtered images. On the other hand, in the Konni dataset, major contribution is from the inverted image predictions. For an elephant warning system, it is important to keep a very low false alarm (false positive) rate. As given in Table 3, the ensemble model produces false positive detections in a total of 701 images in the dataset. Since false positive detections in temporally adjacent frames of a video are unlikely, it is possible to reduce the false alarm rate further by taking a combined decision from the frames in a temporal window. We have also evaluated the feasibility of deploying the proposed ensemble model on an edge device for real-time elephant monitoring. For this, we used an NVIDIA Jetson Xavier NX module with a YOLOv8m model with TensorRT INT8 quantization. We achieved an inference speed of about 6 frames per second, which is reasonable for real-time applications.

5. Conclusions

In this paper, we proposed *InfraNet*, a thermal image object detection method for real-time wildlife monitoring. The proposed ensemble model combines predictions obtained from multiple preprocessed image variants using the pre-trained YOLOv8m model. The experimental results show significant improvements in recall performance without affecting the precision on the considered datasets. We have also introduced a large-scale annotated thermal image dataset of elephants, human, and common domestic animals. The feasibility of real-time deployment of the *InfraNet* model on edge devices is verified using an NVIDIA Jetson Xavier NX module.

Acknowledgment

This work was funded and supported by iHub DivyaSampark, Indian Institute of Technology Roorkee.

References

[1] Infranet elephant thermal dataset. `https://tinyurl.com/infraredElephant`.

[2] Chandan, M. Chakraborty, Anchal, et al. *IEEE Sensors Letters*, 8(9):1–4, 2024.

[3] FLIR Systems, Inc. FLIR ADAS Thermal Dataset. `https://www.flir.com/oem/adas/adas-dataset-form/`, 2018. Accessed: Oct. 28, 2023.

[4] A. G. Hart et al. Can handheld thermal imaging technology improve detection of poachers in african bushveldt? *PLOS ONE*, 10(6):1–13, 06 2015.

[5] S. Hwang, J. Park, et al. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of IEEE Conference on CVPR*, 2015.

[6] X. Jia, Zhu, et al. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF ICCV*, pages 3496–3504, 2021.

[7] G. Jocher, J. Qiu, and A. Chaurasia. Ultralytics yolo, Jan. 2023.

[8] T. Lin, M. Maire, S. J. Belongie, et al. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[9] H. Lv, P. Shan, H. Shi, et al. *Signal Processing: Image Communication*, 102:104703, 2022.

[10] Y. Oishi et al. Animal detection using thermal images and its required observation conditions. *Remote Sensing*, 10(7), 2018.

Location	Class	Baseline		Inverted		Bilateral		Inv+Bilateral		Ensemble	
		P	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
Trivandrum	Elephant	–	–	–	–	–	–	–	–	–	–
Trivandrum	Person	0.372	0.866	0.317	0.958	0.710	0.783	0.508	0.924	0.263	**0.978**
Trivandrum	Other	0.999	0.534	0.999	0.721	0.998	0.796	0.998	0.908	0.997	**0.930**
Konni	Elephant	0.977	0.438	0.976	0.508	0.968	0.354	0.977	0.431	0.959	**0.573**
Konni	Person	0.992	0.860	0.977	0.88	0.966	0.854	0.916	0.892	0.912	**0.920**
Konni	Other	–	–	–	–	–	–	–	–	–	–
Chilla	Elephant	0.993	0.124	0.988	0.318	0.982	0.299	0.979	0.616	0.979	**0.767**
Chilla	Person	0.278	0.250	0.326	0.272	0.204	0.277	0.328	0.594	0.249	**0.702**
Chilla	Other	–	–	–	–	–	–	–	–	–	–
Overall	Elephant	0.967	0.352	0.960	0.456	0.963	0.339	0.960	0.481	0.936	**0.625**
Overall	Person	0.925	0.845	0.899	0.867	0.927	0.837	0.867	0.886	0.805	**0.917**
Overall	Other	0.460	0.534	0.577	0.721	0.527	0.796	0.651	0.908	0.490	**0.930**

Table 2: Frame-level precision and recall performance

Location	Class	Baseline		Inverted		Bilateral		Inv+Bilateral		Ensemble	
		TP	**FP**	**TP**	**FP**	**TP**	**FP**	**TP**	**FP**	**TP**	**FP**
Trivandrum	Elephant	–	87	–	67	–	–	–	50	0	204
Trivandrum	Person	78	553	367	1245	46	37	292	356	783	2191
Trivandrum	Other	108	3	825	2	890	5	4071	9	5894	19
Konni	Elephant	2422	85	3940	130	826	110	2165	74	9353	399
Konni	Person	2294	38	4476	91	5183	306	10614	1735	22767	2170
Konni	Other	–	1828	–	1620	–	1238	–	706	0	5392
Chilla	Elephant	108	2	731	2	807	21	3032	73	4678	98
Chilla	Person	50	192	50	132	66	520	301	566	467	1410
Chilla	Other	–	78	–	131	3	208	2	312	5	729
Overall	Elephant	2530	174	4671	199	1633	131	5197	197	14031	701
Overall	Person	2422	783	4893	1468	5295	863	11207	2657	24017	5861
Overall	Other	108	1909	825	1753	893	1451	4073	1027	5899	6140

Table 3: Frame-level performance in terms of true positive and false positive detections

[11] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[12] H. Ritchie. The state of the world's elephant populations. `https://ourworldindata.org/elephant-populations`, 2024.

[13] A. P. Sayakkara et al. Eloc: Locating wild elephants using low-cost infrasonic detectors. In *2017 13th International Conference on DCOSS*, pages 44–52, 2017.

[14] A. Szenicer, Reinwald, et al. *Remote Sensing in Ecology and Conservation*, 8(2):236–250, 2022.

[15] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *6th ICCV (IEEE Cat. No.98CH36271)*, pages 839–846, 1998.

[16] J. Zhou, B. Zhang, et al. Yolo-cir. *Infrared Physics Technology*, 131:104703, 2023.