SMALL BIRD DETECTION USING MARKOV RANDOM FIELDS AND YOLOV8

A Thesis

Presented to the

Faculty of

California State Polytechnic University, Pomona

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

In

Computer Science

By

Stuti Trivedi

2024

THESIS: SMALL BIRD DETECTION USING MARKOV RANDOM FIELDS AND YOLOV8

AUTHOR: Stuti Jayeshkumar Trivedi

DATE SUBMITTED: Spring 2024

Department of Computer Science

Dr. John Korah

Thesis Committee Chair

Assistant Professor of Computer Science

_____

Dr. Markus Eger

Assistant Professor of Computer Science

_____

Dr. Janel L Ortiz

Assistant Professor of Biological Science

_____

ABSTRACT

Camera-trap brings revolution in Wildlife research and observation from last decade, its popularity increased as Machine learning and computer vision gets outstanding results for wildlife animal detection. Numerous benefits aligned with this combination, which starts from wildlife observation, ecology, and biodiversity research, promoting wildlife conservation, forest restoration, improving law and enforcement in Illegal Poaching activities to human-wildlife conflict mitigation. Animal detection becomes possible because of object detection, which is a well-known concept of Computer Vision that includes classification and localization of objects from given instances. Existing models easily identify large animals and objects from images, currently it is in demand to detect small and distant birds. In this research, we explored a non-neural network-based Markov random field technique and compared it with neural network-based You Only Look Once version 8 (YOLOv8) algorithm to identify distant birds.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1 : INTRODUCTION

In 1956, the camera-trap was introduced and in 1995 using the formal mark and recapture model (Gysel, 1956) (Karanth, 1995) was introduced to observer population of, whereby marking on animal researchers count population of wildlife. Karanth mentioned its usefulness for population ecology by identifying Panthera Tigris in Nagarahole, India (Schneider, 2018). Camera traps, activated by heat or motion, have played a pivotal role in ecological research for over a century, undergoing a significant transformation in recent times. Such motion activated devices are helpful to observe wildlife activities without human interaction. In camera traps data collection is done by less disturbance to wildlife entities. Motion triggered camera-traps are more useful due to its commercial availability and wide range of sectors aligned with it. Observing wildlife activities, animal habitat, lifestyle, and reducing over – exploitation of natural and environmental resources, camera traps have been used for decades, it is also helpful to count forest animal populations and provide valuable data for management decisions.

To reduce and track wildlife poaching, animal detection helps law enforcement to detect and prevent wildlife trafficking and poaching activities. It also helps to identify and detect small and rare animals. For the research, there are plenty of references available which discuss wildlife animal monitoring, like radio tracking, wireless sensor network tracking, satellite, and global positioning system (GPS) tracking (Godley, 2008) (Hulbert, 2001), and monitoring by motion sensitive camera traps. It also helps confirm animal wellbeing in tourist areas of the forest and in the zoo and animal detection in forest areas can help to see animal and environment relationships. It is helpful to view the ecology and animal lifestyle, simultaneously protecting the animal and

wildlife resources from illegal activities such as habitat destruction, illegal fishing, and illegal trade of endangered species.

Object detection has been employed for identifying objects in an image or video frame for many years, where the image is not only limited to recognition of the object but also presents the boundaries with bounding boxes. Object detection detects and localizes the objects which belong to various classes. The main use case of object detection is in autonomous vehicles, surveillance systems, robotics, and image retrieval fields. Objects can be defined by their shape, texture, color, and other traits identical in most cases. Once the camera-trap sensors capture the target image then it indicates its position in the given image, whereas for video sequence object detection tracks the position, size and pixel location for objects seen in various frames.

Detection of an object is a classification problem, as it tells whether a specific object is present or not in the given image. The object detection pipeline is defined with three stages: (I) Region Selection the image with multiscale sliding window(Lee, 2017), which allows detection of all possible positions of objects with various sizes, (II) Feature Extraction: it is the process of transforming input image to features, which are informative and relevant to the problem statement such as, color texture, edges, which help to separate different objects and determine their relationship. These features are input for the Machine Learning (ML) and Deep Learning (DL) model to train on different patterns and relationships between objects, and (III) Classification: is a type of supervised learning where the goal is to predict the categorical label of given input data. In a classification task, the machine learning model is trained on a dataset that includes input-output pairs. The ML and DL models are trained on various dataset with labeled image examples.

Animal detection and tracking has been more at the center of research in the last few years as Computer Vision and Deep Learning offers more efficient techniques. Due to fast and accurate detection capacity You Only Look Once (YOLO) (Huang, 2018) became famous, whereas the other latest versions like YOLO (Liu, 2018), YOLOv3, YOLOv5 (Li, 2021) made sustainable enhancements to the original algorithm. YOLOv3 advanced the multi-scale prediction of classes, and improved detection performance. YOLOv5 boasts more simplified architecture that results in faster detection with improved accuracy.

For insights into wildlife activities, camera trap and object detection combination show outstanding results in research where traditionally non-neural network-based approaches were used for object detection in edge computing and Internet of Things devices which rely on hand-engineered features and heuristic to detect objects, on the other side neural network-based approaches can automatically train itself for complex patterns and various features. The current model easily identifies large animals, humans, vehicles, and birds, but it struggles to detect distant and small birds. Moreover, sometimes the captured image color contrast makes it harder to identify the animal/bird whose color is the same as background or foreground. The small objects have ambiguous boundaries and low resolution, hence small object identification is challenging. Existing models are based on high-level Convolutional Neural Networks (CNN) features, which require high performance computational power due to depth of the neural network. Current demand is to make an efficient model which can easily identify the small and distant bird.

The non neural network-based Markov random field is one of a statistical and probabilistic model which is used to represent contextual relationship on the lattice or graph structure. The main concept is value of a pixel in an image depends primarily on its immediate neighbor pixels. By

focusing on all pixels of the image MRF analyzes all pixel structures. For an instance in an image with unwanted noise MRF can be used to smooth the image by considering the neighboring pixel values. Similarly, this research tries to utilize the main concept of analyzing neighboring pixels and then label the pixel as if it is part of the object or not. For this task this research will explore Markov model, with Markov Random Fields and Hidden Markov Model. On the other side, the Neural network based YOLOv8 algorithm is a base model for this research to compare the results and outcome. The YOLOv8 build upon You Only Looks Once series which is well known for its speed and accuracy in detecting objects. YOLOv8 is the best fit for real time applications such as autonomous driving, surveillance, Retails and Healthcare industry. Both concepts of MRF and YOLOv8 are capable of handling dense pixel clusters and abstract meaningful insights from image. This research will explain and compare these two methodologies.

MRF models the special dependencies and relationships between different parts of an image, where it models the pixel structure and compare pixel with the neighboring pixel, as per the comparison it identifies whether a given pixel is part of object or not, also it looks for previous patterns of neighboring pixels special dependences and take decisions as per the analysis. The challenging part in Object detection is object boundary identification, where the Markov random field performs better without neural network or model training. In MRF every pixel value is compared to its neighboring pixel value with prior (previous pixel cluster) knowledge and then assign a label to the pixel, whether it is part of an image or not. Here, this method utilizes previous pixel cluster and its label values to identify whether the next pixel is part of an object or not and detect the object boundary and identify the detected object with label assigned to it. This process of detecting and labeling all pixels makes MRF a suitable choice to detect a small bird in a dense

and distant image. This research compares results from Neural network based MRF and Non-neural network based YOLOv8.

The primary objective of this subsection is to provide a roadmap of the thesis report. It first presents various related work on non-neural network-based object detection methods and Neural network-based object detection methods in Chapter 2. It elaborates on research challenges, research questions and objectives in Chapter 3. This research mentioned three different techniques: Markov random Fields technique, YOLOv8 and absolute difference, and compared and analyzed the results. Technical Preliminaries are mentioned in Chapter 4 with detailed description and mathematical concept of Markov Random Fields (MRF), Gaussian and Morphological filters. Chapter 5 has the proposed methods. Chapter 6 mentioned experimental setup with software and hardware configuration, utilized dataset for this research with hypothesis with analyzed result. Chapter 7 has conclusion and future work.

CHAPTER 2: LITERATURE SURVEY

This chapter aims to understand object detection evolution in-depth with starting from non-neural network based to neural network-based approach. Each method offers unique advantages and challenges, which are critical in understanding their applicability across various real-world scenarios. Non-Neural Network-Based Approach for Object Detection: Histogram of Oriented Gradients (HOG) (Nguyen, 2019) and Support Vector Machines (SVM), this classical computer vision approach relies on extracting features based on gradients in image intensity. Firstly, HOG captures the local object shape information and SVMs are responsible to classify those descriptors into classes. These methods are effective for some cases, however HOG + SVM struggles with diverse object variety. To detect small targets in infrared image sequences author Xinyu Wang introduced the gray-scale morphology (Wang, 2009) for the removal of large image background regions which is working on fast top-hat transformation.

For medical research Markov Random Field (MRF) showed outstanding results, one of such research by Marroquin et al. (Marroquin, 2002) was performed on Magnetic resonance imaging (MRI), where MRFs were used to segment an MRI image into disjoint of regions. The main idea of the paper is using Bayesian Estimation Theory and MRF to segment a slice of MRI (Geman, 1984).

In research work by Benedek et al (Benedek, 2007) proposed a three-layer Markov Random field structure, where each layer focuses on different information to detect small objects in continuous timestamped frames of video. The first layer deals with direct pixel information, and the second layer will manage correlation features on pixel information taken from layer one.

The last layer integrates both layer details and produces a final segmentation that distinguishes between static objects and moving objects.

Neural Network-Based Approach for Object Detection are Region-Based Convolutional Neural Networks (R-CNN) (Christin, 2021) and Its Variants like R-CNN, Fast R-CNN (Girshick, 2014), and Faster R-CNN (Girshick, 2015), revolutionized object detection history. R-CNN generates a region proposal (object region) using a selective search algorithm, to extract features CNN is applied and then to classify the bounding boxes to objects, and each object requires separate training. Whereas Fast- RCNN (Region Based Convolutional Neural Networks) works with a region of Interest (ROI) pooling layer, where it doesn't process a region separately, instead it passes entire image in a single forward pass through the CNN, then it generates feature vectors for a region by applying the feature maps which simplifies the training process compared to R-CNN. Instead of using selective search, Faster R-CNN works with Region Proposal Network (RPN) which is a fully convolutional network that uses the convolution feature with CNN. The end-to-end object detection without explicitly specifically defining the features, which make it best fit for the different object types and scenes.

The availability of open-source detection models offers an opportunity for individuals to independently assess and apply pre-trained tools. A notable example is MegaDetector, a system-agnostic object detection model developed by Microsoft explicitly for processing five camera trap data, the neural network-based object detection model (Beery, 2019). This freely accessible model, trained on a vast dataset comprising millions of global images, is proficient in identifying three object classes in images: humans, animals, and vehicles. Consequently, it implicitly recognizes images without any objects from these classes. The automated categorization of images into these

classes holds the potential for significantly faster processing compared to manual human efforts, with speed limitations primarily dictated by computer processing capabilities (Beery S, 2020).

From last few decades agriculture, construction, public safety as well as forestry industries started using Unmanned Aerial Vehicle (UAV) prepared with drones, which are not only able to capturing images but also able to identify objects using Artificial Intelligence (AI) and Computer Vision. In such a scenario, detection of small objects from aerial images becomes a challenging task, including background complexity, resolution changes and limited range of pixel representation. To address these issues, recently Xu Yan (Yan, 2023) proposed two methods (I) Magnifier Method where feature extraction in YOLOv5 is enlarging small bird image. Firstly, crop the image from 1920 x 1080 into 2 rows and 3 columns, resulting in 6 small images with 640 x 640 pixels each. All these small images were trained into Yolov5 and after detecting the object all images were merged. To remove redundant bounding boxes Non-Max Suppression (refine the object detection process by eliminating overlapping bounding boxes) was applied to resultant images. (II) Reconstruction Detection Branch add a small object detection layer to extract features of small objects in YOLov5 to improve overall result. RCD method extract 160 * 160 feature map from the 640 * 640 input image by allowing models to detect 4 * 4-pixel objects. This may also lead to a slight reduction in precision for larger objects as they truncated during the mage cropping processes.

In the paper written by Yan et al. (Yan, 2023), YOLO-inspection architecture was introduced on top of the Darknet53 (Yan, 2023) (Mahum, 2022) in YOLOv3. This architecture processes inspection-like module before the information fusion stage. The model uses convolution kernels of 1 x 1 and 7x7 size. The 1x1 integrates cross-channel information and changed the

numbers of channels and 7x7 kernels provides semantic information for small objects. This method achieves mAP (mean average precision) 78.37%, results in better performance than YOLOv3. The paper focuses on PASCAL VOC dataset only, it did not discuss the effect of wide range of dataset on the model. Moreover, the paper did not discuss details on selection of hyperparameters, like ratio of channels in the high-level and low-level layers.

The paper by Kim et el. (Kim, 2022) discussed about head and attention, where they proposed a modify PANet structure (an architecture that aggregates various levels feature maps) to enhance information about small objects in feature map by using additional feature heads to improve detection performance for small objects by connecting heads to law-resolution feature maps. In VisDrone 2020-DET dataset the researchers get 4.1% improvement in mAP. The modified CBAM enhances information about small objects in PANet, and backbone network which is admirable. The channel attention module of the CBAM is modified and added to the PANet. This modified CBAM enhances information about small objects in the PANet, backbone network, and bottom and top layer feature maps, especially in P2 and P3 layers.

To improve accuracy and speed YOLOv5 was introduced by (Li, 2021), which consists of CSPDarknet3 backbone and feature extraction. This paper offers a simple approach to scale model size up or down based on target device computational resources. This paper also introduces data augmentation technique during training phase, to enhance model activity to generalize different scenarios and improve robustness. As it uses data augmentation technique, it requires large and diverse dataset which is one of the limitations. Moreover, it uses a single scale interface strategy, which could be an obstacle to detecting small birds.

In the work by Li et al (Li, 2020) researchers were trying to detect table tennis balls, for that they use CNN with residual connections to extract image feature. This architecture helps in capturing complex patterns in images. This method also constructed feature pyramid, which is admirable, as the feature map with rich semantic information in the upper layer is fused with the layer containing rich object position information in the lower layer. It achieves high detection accuracy of 93% and fast detection speed of 95 frames per second (FPS). All research did not compare their proposed method with other state-of-the-art object detection models. This research will focus on small bird detection with the help of non-neural network based and neural network-based approaches. For non-neural network-based method, it will explore the scope of MRF technique to observe the pixel change and for the neural network-based technique it will perform a well-known YOLOv8 algorithm to the wildlife dataset. In the next chapter the research questions and challenges are mentioned with objectives.

CHAPTER 3: RESEARCH QUESTIONS, CHALLENGES AND OBJECTIVES

In recent years Object detection has achieved more noteworthy results due to the increasing developments in deep learning and neural networks. This progress has achieved high accuracy in identifying and localizing numerous objects of diverse domains, from automation systems to wildlife monitoring. However, despite these successes, it deals with a few challenges, particularly in an over-changing and complicated environment. For example, detecting camouflaged or small birds in natural set up remains difficult due to occlusions, shadow change, variable lighting, and highly cluttered background.

3.1 Research Questions

This research will answer following two questions by comparing Markov Random Filed and YOLOv8. How do we differentiate birds from similar-sized leaves in images, especially when the leaves' motion could mimic bird activity? Such critical scenarios arise when complexity of natural background is high as leaves and birds may share same size and colors, and leaves movement will be similar with bird activity. Addressing this challenge will require analyzing not just the static behavior but also dynamic behavior with image context. Techniques like Markov random field and observing pixel changes over consecutive images will help to record all information of the image, such methods could improve the accuracy of differentiating between birds and leaves.

How effectively does non-neural network-based Markov Random Field technique identify small animals in dense and blurring background images? In image processing and pattern recognition, accurate image segmentation, classification and object region detection is crucial. A

Markov random field shows the relationships between various parts of the image, making it possible to incorporate contextual information about neighboring pixels or regions. This is beneficial for distinguishing objects from their backgrounds based on the properties of local image areas.

3.2 Research Challenges

Despite the potential of object detection in the field of neural network and non-neural network-based approaches, implementing these strategies faces unique challenges. Inadequate detection of small and distant birds due to their size, blending with background and low resolution. Due to very few pixels availability to create bird in image, it reduces detailed information to detect and classify object, which increased challenges in differentiating these birds from background noise. In addition, as the distance between camera set-up increased, object resolution increases, which makes detection process more challenging due to lack of picture clarity.

In response to these research challenges, the primary objective of this report is to evaluate Markov Radom Field (MRF) techniques in detecting small objects within dense environments, whether MRF handles spatial dependencies against noisy background or not. On the other side YOLOv8's performance will be evaluated, and finally comparative analysis will be conducted to observe the results and provide detailed performance evaluation into which methods handled small bird.

3.3 Research Objectives

To address these challenges, this research will compare both MRF and CNN method under given dataset to compare in terms of Precision, Recall, and computational time. This

comparison will handle different lighting conditions, small sized bird, and background complexity. Moreover, it will mention the advantages and disadvantages of this method over small birds. The next chapter represents the technical background of Markov Random Field.

CHAPTER 4: TECHNICAL PRELIMINARIES

This chapter describes the fundamental and statistical concept of Markov model, a class of probabilistic frameworks for modeling random systems that change over time. It starts with the Markov model concept, which elaborates from Markov property. Simple Markov Chain and Hidden Markov Model represents a more complex extension of Markov Chains where the states are not directly observed. Instead, the model includes observable outputs that depend probabilistically on the hidden states.

Markov chain (Blake, 2011) mathematical formulas are useful in image segmentation, which divides an image into multiple segments or set of pixels, known as objects. Image segmentation is the process of transferring an image into segments where each segment represents a meaningful area. To locate objects and boundaries such as lines and curves, image segmentation is the best choice in the image processing field. It is useful in various applications such as medical imagining, video surveillance, machine inspection and in autonomous driving. For instance, for segmentation of tissues or detection of anomalies over large areas of image.

An image can be represented as a collection of nodes, which corresponds to individual pixels or group of pixels. A graph G = (V, E), is combination of Vertices V and Edge E, where V = (1, 2, . . . , *i*, . . . , N) corresponds to pixels of image and E = edge (*i*, *j*), where *i*, *j* ∈ V, the edges are undirected means *(i, j)* and (*j*, *i*) refers to same edge. The collection of pixels of image, which create node is known as superpixels, and a pair of superpixels creates an edge if the superpixels share a common boundary (Blake, 2011). In the Markov model these superpixels are the deciding factor in object region detection.

4.1 Markov model

Markov models describe a system of interconnected variables models show associations between a few pairs of pixels, these pixels are defined neighborhood pixels, as they share an edge E. If one has values of neighbors, that means one has all necessary and essential information required to determine the probability of variables values. We can relate this in our houses and neighborhoods as what happens in one house directly affects the house next to it, not the whole town. The main attraction of Markov model is Knock-on-effect (Blake, 2011) which explicit short-range linkages give rise to implied long-range correlation. It says that while model explicitly shows short range interaction between short range pixels, these interactions imply longer-range correlations through a chain of local interaction. This outstanding strength of Markov model allows the model to cover border area of image and correlation within the data without direct computational expense of managing long-range dependencies explicitly (Blake, 2011).

4.1.1 Markov Chain

The simplest Markov model known as Markov chain is a continuous sequence of random variables $X = X_1 + X_2 + \cdots N$ which has a joint distribution mentioned as $P(X_i \mid X_{i-1}, X_{i-2}, \dots, X_1)$. The basic example of Markov Chain is Whether forecasting, that the probability of today's whether is explicitly connected with yesterday's whether and implicitly linked as a knock-on effect to all previous days. The first order Markov model explicitly includes longer range dependencies similarly the three successive days conditional dependencies will achieve by multiplying matrices for two successive days. The Markov model can be represented as directed or undirected graph (Figure 1), in 2D images, model expressed in one form can be

expressed in other forms. However, many Computer vision models are expressed as undirected graph hence this report will use undirected graph for Markov model evaluation.

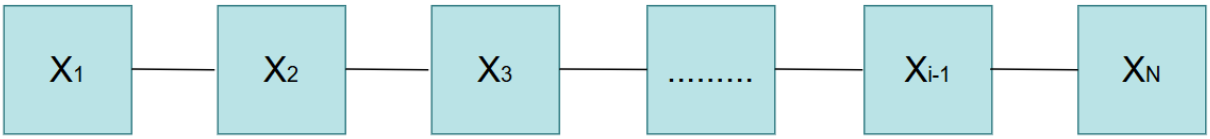


Figure 1 Undirected Markov model

The Markov models show their importance as a prior model for state variables Xi that will be estimated from corresponding observations $Z = (z_1, z_{2,\ldots,} z_i, \ldots, z_N)$ Observation is an instance of a random variable X. The observations z are themselves considered to be instantiations of a random variable Z representing the full space of observation. This arrangement creates an inferenced challenge where the posterior distribution of the potential state X, with observation Z is calculated by Bayse's formula. (Blake, 2011)

The posterior distribution (Blake, 2011) represents the updated probability of the model's state given the observed data. This distribution is central to Bayesian inference, where we start with a prior belief about the distribution of states (prior distribution) and update this belief based on new, observed data (likelihood) to obtain the posterior distribution. The Bayesian method gives a probabilistic framework for making inferences about the image based on observed pixel data and prior knowledge.

In image segmentation, bayes formula will classify each superpixels into different segments based on image characteristics such as color, texture, and intensity, here each label corresponds to a different segment of image.

$$P(X = x \mid Z = z) \propto P(Z = z \mid X = x)P(X = x) \quad (1.1)$$

Where, *P (X = x)* represents the prior distribution over states, indicating what is known about the states X before any observations are made.

$$P(x \mid z) \propto P(z \mid x)P(x) \quad (1.2)$$

constant of proportionality would be fixed to ensure that $\sum_x P(x \mid z) = 1$, this is denoted as

$$P(X \mid Z, \omega) \propto P(Z \mid X, \omega)P(X \mid \omega) \quad (1.3)$$

Where $\omega \in \Omega$ prior model or observation model or both, and Ω is the Model space. Where the model space refers to all states models can assume based on their parameters, either the prior model or observation mode. Prior model is a part of probabilistic model which represents what is known or assumed about the system before new observation is added. As per Bayesian terms, this is the "prior distribution," which encodes any pre-existing beliefs about the parameters or states of the model. And the observation model relates the observable data to the states of the model. It often specifies how likely it is to observe the data given states or configurations of the model.

4.1.2 Hidden Markov Model

Hidden Markov Model is itself represented as a Markov chain, which in the first-order caise was decomposed as a product of conditional distributions. $P(z \mid x)$ implies probability of observations. The observation at site *i* depends only on the corresponding state. (Blake, 2011). HMM represents systems which follow a Markov process with hidden states, here hidden states mean a previous observation which impact on the current stage. HMMs are widely used in biometrics, speech recognition and finance due to their capability to observe time series data.

The key concept of HMM is hidden state and observable state. The hidden state is not directly observable, however it impacts the outcome, whereas the observable state is a state that is seen in terms of output. HMM works on Markov property: the probability of moving to the next state depends only on the current state. Below is the probability explanation of HMM.

In other words: $P(z \mid x) = P(z_N \mid x_N)P(z_{N-1} \mid x_{N-1}) \cdots P(z_1 \mid x_1)$ (1.4)

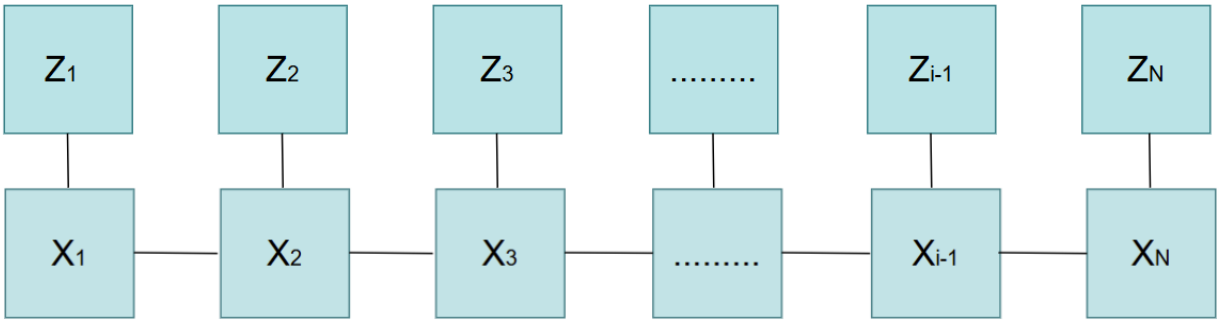


Figure 2 Undirected Hidden Markov Model

$$P(z \mid x) = \Phi_N(x_N)\,\Phi_{N-1}(x_{N-1}) \cdots \Phi_1(x_1) \quad (1.5)$$

where trivially $\Phi_i(x_i) = P(z_i \mid x_i)$

Now the posterior for the state x of the pixels is obtained from Bayes's formula:

$$P(x \mid z, \omega) \propto P(z \mid x, \omega)P(x \mid \omega) \quad (1.6)$$

HMMs treat the pixels or regions of an image as observable states, while the hidden states represent the underlying object categories or features that are not directly visible. The process begins by initializing the model with prior knowledge about the probable distribution of hidden states and the relationship between these states and the observed data. As the model scans through the image, it uses the observation likelihood $P(z_i \mid x_i)$ to calculate the probability of the current

observed data given the hidden states, which might correspond to various object parts or characteristics. Through the application of Bayes' formula, the model updates the posterior probability *P(x/z,ω),* combining this likelihood with prior state probabilities to refine its predictions.

$$P(x \mid z, \omega) = \frac{1}{Z(z,\omega)} \exp\big(-E(x, z, \omega)\big) \tag{1.7}$$

where $E(x, z, \omega) = \sum_{c \in C} \Psi_c(x, \omega) + \sum_i \Phi_i(x_i, z_i)$

The energy functions for many commonly used Markov models are written as a sum of unary and pairwise terms,

$$\mathrm{E}(x, z, \omega) = \sum_{i \in V} \Phi_i(x_i, z_i, \omega) + \sum_{(\mathrm{i,j}) \in \mathrm{E}} \Psi_{\mathrm{ij}}\big(\mathrm{x_i}, \mathrm{x_j}, \omega\big) \tag{1.8}$$

HMM uses forward and Viterbi algorithm (Ardo, 2007) for interference, where the forward algorithm (Yu, 2006) is used to calculate the probability of an observed sequence given the model parameters. This algorithm employs dynamic programming to efficiently compute the likelihood of the observation sequence by breaking the problem into small parts. When a task requires precise state estimation over time, the Viterbi Algorithm determines the most likely sequence of hidden states. In contrast, the Forward Algorithm calculates the likelihood of an observed sequence, providing a measure of how well the model fits the data.

This posterior probability is expressed as an energy function, which incorporates both unary potentials (the likelihood of each state given the observation) and pairwise potentials (the spatial relationship between states). By minimizing this energy function, the HMM can infer the most probable hidden states, effectively identifying and segmenting objects within the image

based on observed pixel data and learned model parameters. This approach allows HMMs to provide a robust framework for object detection by capturing both local and global contextual information within the image. The next chapter discussed the proposed methodologies with details of MRF technique associated with this research.

CHAPTER 5: METHODOLOGY

This chapter first discusses the Markov random Field (MRF) technique (Ait-Aoudia, 2011) (Geman, 1984) (Cross, 1983) and then explores three consecutive image pixel changes and at least it displays potential of YOLOv8 to detect small birds. By analyzing results with all these three methodologies it will choose the one best for the given setup of research.

5.1 MRF

Among all non-neural network base object detection methods Markov Random Filed is the method which utilized contextual relationships between different elements within an image, such as pixels or regions. This capability allows pixels to effectively model the spatial dependencies and inherent structure of the visual data, which is crucial for accurately segmenting and identifying objects in complex scenes. The method uses various prior knowledge and constraints about the target application which other non-neural network methods like HOG, background subtraction and template matching are not able to provide. This adaptability makes MRF suitable for a wide range of applications, from medical imaging to remote sensing, where precise modeling of spatial relationships is vital.

This research will explore MRF method as per given equation (1.8) with unary and pairwise potentials, where unary potential will observe neighbor pixels and pairwise potential will observe hidden variable (hidden observation). (Blake, 2011)

$$E(x, z, \omega) = \sum_{i \in V} \Phi_i(x_i, z_i, \omega) + \sum_{(i,j) \in E} \Psi_{ij}(x_i, x_j, \omega) \quad (1.8)$$

This equation has two elements where Unary Potential (first element) depends only on a single variable, which depicts a location of image at a particular location or pixel color. Z_i is measurable attribute in present of observed pixel values x_i, where ω represents presence of model parameters, means how observed data influenced the state x_i, and then as per the information it will adjust sensitivity of model to differentiate between observed data and expected states based on the model.

Unary potentials evaluate the likelihood of appointing a certain state or label (edges, object boundary or corner) to each pixel or region based on local image data of that pixel. Pairwise Potential (second element) will evaluate dependency between x_i, and x_j, neighboring pixels in the image, and it will consider how two adjacent pixels share the same label by being part of the same object. Pairwise potentials are important for ensuring consistency in labeling across the image.

As image is a collection of different pixel values, where each pixel is responsible to represent the RGB color space values of that region. Observed values for each pixel i included various features that describes pixel properties, such values are gradient magnitude equation (1.9) and gradient direction equation (1.10), which is strength of gradient at pixel i, which indicates how much the intensity of the image changes at that point. Gradient direction is orientation of gradient at direction i, which shows direction at which pixel intensity changes the most.

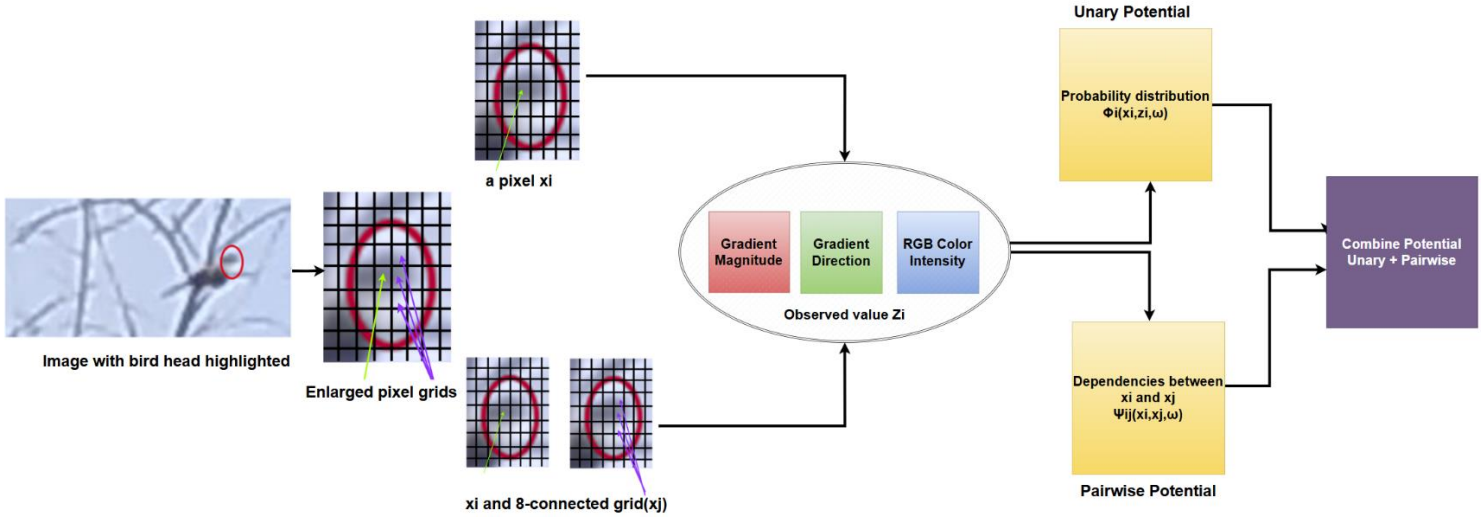


Figure 3 MRF architecture

$$\text{Gradient Magnitude} = \sqrt{\{G_x{}^2 + G_y{}^2\}} \tag{1.9}$$

$$\text{Gradient Direction} = \arctan\left(\frac{G_y}{G_x}\right) \tag{1.10}$$

Where G_x and G_y are the gradients in the *x* and *y* directions respectively.

Extract the RGB values, Gradient Magnitude, Gradient Direction of each x_i decides the likelihood of pixel belonging. For each possible label x_i, which is part of image compute the likelihood of *i* belongs to that label (bird or background), with respect to its observed values Z_i is decided with probability distribution as per equation (1.5). These features are defining factors for unary potentials which measure likelihood of each pixel being assigned a particular label.

The pairwise potentials focus on the relationship between pair of neighboring pixels rather than individual pixel characteristics. Firstly the neighboring pixel (*i*, *j*) is determined with an 8-connected grid, for each pair of neighboring pixels (*i*,*j*), extract the relevant features from both

pixels, where features are color, intensity, gradient direction and gradient magnitude. After this stage model dependency is determined to define how neighboring pixels share the same label or have different labels based on their observed features, hence pairwise potential is a key factor to decide that current pixel have same property as the 8-connected neighbor, if yes, then it counts that pixel as part of label (object either bird or background). Calculate the pairwise potential for each pair of neighboring pixels. This potential penalizes label assignments that are inconsistent with the observed relationship between the pixels.

Unary potentials evaluate the likelihood of assigning a particular state or label to each pixel based on local image data, such as pixel intensity or color, while pairwise potentials ensure consistency by considering the dependencies between neighboring pixels. Finally, the energy function equation (1.8) is minimized to achieve optimal labeling across the image. The final step includes post-processing techniques to refine the detected objects, ensuring accurate and coherent segmentation of objects from the given image.

Overall, the energy function will find the most probable state configuration for the whole image and segment the object from the image. The output is an image where each pixel is assigned a label that collectively minimizes the energy function. The labeling not only fits the observed data per pixel but also maintains consistency between adjacent pixels, resulting in detected objects region in the image. In the next subsection, this report mentioned another method to compare with this MRF technique.

5.2 Absolute pixel difference

While researching and experimenting on Markov Random field, this research also observed continuous images of camera-trap image captured time. Such three consecutive images

will first pass on gaussian blur filter and histogram equalization, then calculate absolute pixel change difference between three images and at last pass morphological filter.

The Gaussian blur (Mishra, 2014) technique will create a convolution mask, which will apply to all those three images, and give more weight to central pixels and less to pixels farther way. This will smooth the image. Gaussian blur is applied using a convolution kernel (also known as a mask or window) that corresponds to the values of a Gaussian function. This kernel moves over every pixel of the image. Each pixel and its neighbors will apply matrix multiplication by the Gaussian kernel, where pixels closer to the center have higher weights due to the properties of the Gaussian function (it falls off exponentially from the center). The output pixel value is the weighted average of the pixel values in the neighborhood defined by the kernel. This blending effect smooths sharp edges and reduces the variation in pixel values, which effectively blurs the image.

Due to lightning and environment condition, improve image contrast by Histogram Equalization (Chauhan, 2011) which will spread out most frequent intensity values, which enhance overall contrast of image. First, the histogram of the image is derived which shows the distribution of pixel intensities in the image. Each bin of the histogram represents the frequency of a particular intensity value. The cumulative distribution function derived from the histogram which is used to map the old intensity values of the pixels to new values that spread more evenly across the available range, thereby improving the image's overall visual quality by enhancing

underrepresented intensities. This process enhances the contrast of areas with subtle intensity variations, making them more distinct.

Finally Morphological filtering (Bethel, 2017) will apply, which is collection of non-linier operations related to the morphology of features in an image. It relies on relative order of pixel values rather than on numerical value. It passes on erosion an operation which shrinks or thins objects in a binary image and removes small-scale noise, a structuring element (kernel) slides over the image and pixels that are completely covered by kernel are retained in resulting

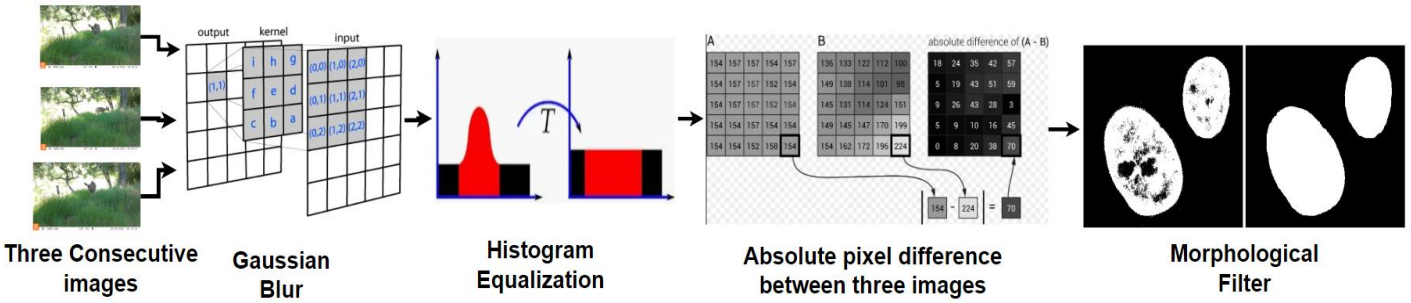


Figure 4 Absolute pixel difference architecture

image. Whereas Dilation will expand or thicken objects in a binary image. And at last Opening operation is an erosion followed by a dilation. It is used to remove small objects from noise. These stages of process will result in proper pixel cluster, enough to decide animal present or not.

5.3 YOLOV8

The standard YOLOv8 is neural network base method that this report will compare with non-neural network based MRF architecture. A simple architecture of You Only Look Once (YOLO) is shown in figure 4. Using a single convolutional neural network (CNN) to simultaneously predict numerous bounding boxes and class probabilities throughout the whole image, the YOLO architecture revolutionizes object detection. By observing the complete image

during training and inference, YOLO attains high speeds and accuracy and gathers contextual knowledge about object classes and their visual appearance. The network creates a grid out of the image, predicts probabilities and bounding boxes for each grid cell. Amongst various versions of YOLO this report focused more on YOLOv8 as its performance and accuracy is high due to latest TPU (Tensor Processing Unit) and it is advanced in varied image qualities, complex scenarios, and better generalization across different environments that were challenging for previous versions.

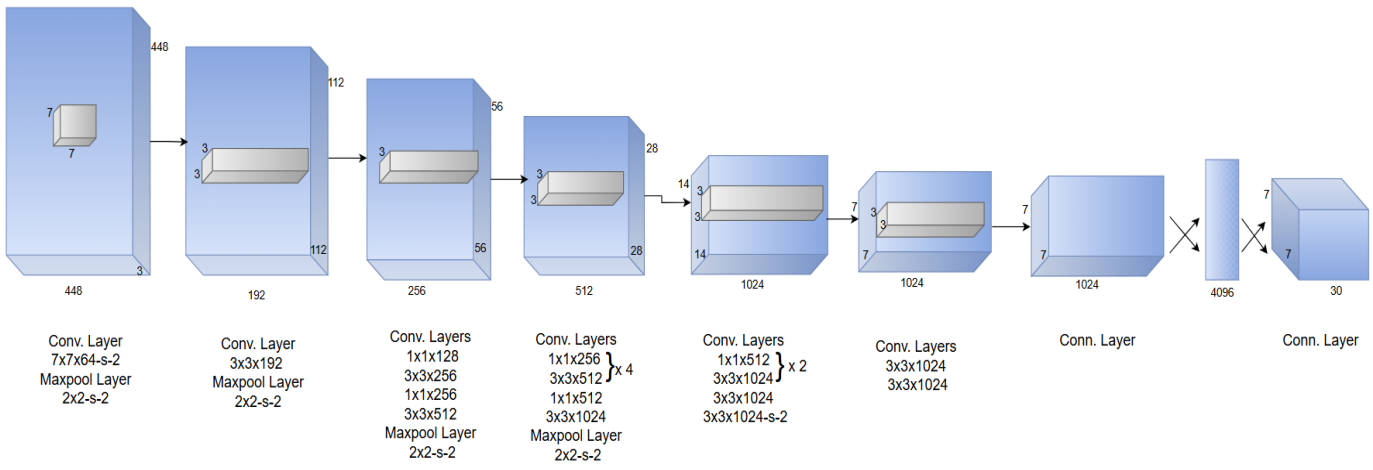


Figure 5 YOLO Farmwork (Redmon, 2016)

YOLOv8 is structured in three main branches where each of the branches performs a series of operations on the given input image. All these will process different scales of the input data for multi-scale feature extraction, which will help to detect small sized objects. In feature extraction, first it splits the input image into three parts with specified axis, each will pass through scaling, modifying, and introducing nonlinear dynamics, these will enhance feature extraction. The split operation will apply different transformations to the same input data. Following a split, the mul operation works with B =2, which specifies the factor by which each element of input tensor (a multi-dimensional data array) will multiple. Next, Add and Power operation where each function

will add biases, it is common function to shift activation function in neural network. Pow = 2 will square the element of tensor, after that the branches multiply all previous result, it is a way to merge features from different layers or transformations. Reshape operation will align the data structure for the next operation. Final concatenation combines all extracted features of object in one structure and make final prediction. Finally, at last it outputs a tensor of shape 1x25200x85, meaning the neural network predicts on a grid of 25200 cells and each cell predicts multiple attributes. (Ultralytics, 2023)

This advanced framework of YOLOv8 makes it versatile in all object detection tasks and makes it able to detect small sized objects as it is utilizing high-resolution feature maps. effectively capturing fine details at various scales, the model can identify small birds within complex backgrounds. In the next chapter, the experimental setup and detailed analysis of the result is described.

CHAPTER 6: EXPERIMENTAL VALIDATION

In this section, the focus will be experimental setup with hardware and software configuration described in subsection 6.1.1 and detailed dataset overview along with types of images and image quality is mentioned in 6.1.2. The hypothesis for the experiment is described in sub section 6.1.3 and data annotation is in 6.1.4. The result analysis factors, and final discussion of the report is detailed in 6.2.

6.1 Experimental Setup

To guarantee repeatability, transparency, replicability, and generalizability, the experimental setup is organized in a particular way. The hardware and software configuration were covered in detail in this subsection. To ensure that the dataset is appropriate for the research objectives, it introduces the dataset utilized in this study and clarifies any adjustments or enhancements that have been made to it. The performance metrics used to measure the effectiveness of our models are model precision, recall, mean average precision and F1 score, in conclusion, the predicted outcomes were compared and analyzed.

6.1.1 Hardware/Software Configuration

The experimental setup configuration for our research consists of both hardware and software components. The hardware configurations are:

- Processor: 11th Gen Intel(R) Core (TM) i5-1135G7 @ 2.40GHz, 1382 MHz, 4 Core(s), 8 Logical Processor(s)
- System Type: x64-based PC

- Experiment performed on Google Collab:
    - Faster GPUs (Graphics Processing Units), which include Nvidia K80s, T4s, P4s, or P100s, depending on availability.
    - 25GB RAM

Our software stack includes programming languages and libraries tailored for machine learning and deep learning tasks, such as Python, TensorFlow, Keras, scikitlearn.

6.1.2 Dataset

The Biological Science Department of Cal Poly Pomona is presently involved in a wildlife animal observation project. Images captured in forest areas, backyards, gardens, farms, human-wildlife interface areas. Camera set up over more than 27 Southern California regions. Images more than 80K, ~70% of camera trap images are empty, due to a high rate of false Positives.

Figure 6 Sample Image with enlarge bird

The dataset consists of 500 images featuring small and distant birds, figure 5 shows sample image. The images vary slightly in size, ranging from 2.06 to 2.17MB, except for those affected by motion blur, where the size ranges from 2.24 to 2.27 MB. All images share the same

dimensions of 3840x2160 pixels, with a horizontal and vertical resolution of 72 dpi. The bit depth for these images is 24, and the resolution unit is set to 2. The images are affected by noise factors such as motion blur and atmospheric conditions.

6.1.3 Hypotheses

Neural network-based methods are superior particularly in complex background and dense images. This research hypotheses that Neural network based YOLOv8 method is superior to non-neural network-based Markov Random Field technique. As neural networks excel in identifying intricate patterns and subtle distinctions in images, which are often challenging for traditional algorithms that do not utilize deep learning techniques.

6.1.4 Data Annotation

Images were labeled using the Computer Vision Annotation Tool (CVAT), categorizing them into two classes: '0' for 'not bird' and '1' for 'bird'. The model training was conducted for 25 epochs with a classification threshold of 0.25 to determine the presence of birds. The training time per image dataset was also recorded to assess the efficiency of the process. This setup enables an in-depth analysis of the model's performance and feature extraction capabilities.

6.2 Result analysis

This section first observes MRF output result, which leads to segmentation and then evaluates results with absolute pixel difference and at last analyzes YOLOv8 result with training graphs.

6.2.1 MRF Results

This section will analyze MRFs proposed method outcome as MRFs model the image as a graph where each pixel (superpixels) interacts with its neighbors. Such framework preserves local pixels details and applies unary potential on it; However, it inherently limits the scope of feature detection because it only considers immediate surroundings. This local focus results in the model missing more extensive or subtle features that don't significantly extract the local features of the image, especially when these features are small relative to the scale of local interactions.

Figure 7 MRF output image

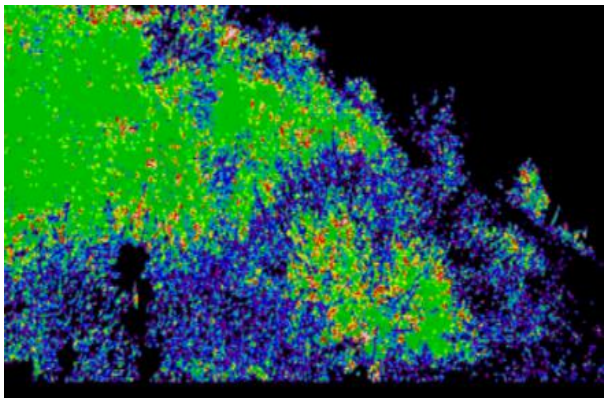


Figure 8 MRF input image

Moreover, the proposed equation struggles to differentiate between actual object and background noise. Also, the previous observation for binary potentials is not more useful in identifying the detailed features and considering the small object as a background. Hence, from the given resultant figure 7, the proposed method equation 1.8 results in a segmentation, where it segments each pixel as per their color and brightness, with weak contextual intersection. This results in the Markov Random Fields technique with proposed equation is best for image segmentation not for object detection.

6.2.2 Absolute pixel difference results

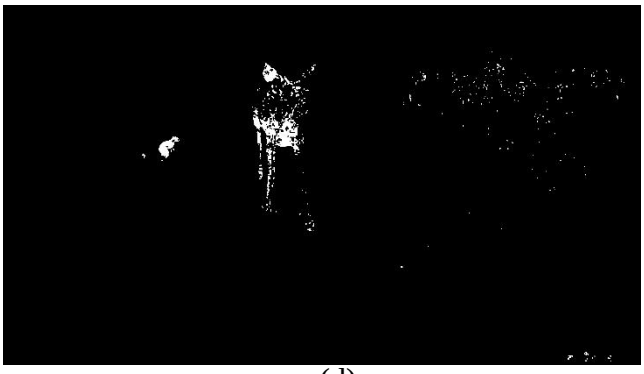
This section observes results by comparing three consecutive images and applying Image processing filters where Gaussian filter improves image clarity by reducing background noise, histogram equalization evenly distributes most frequent intensity values. The morphological filter is a key factor in this absolute difference technique, as it results in clear precise animal image, this became possible due to its erosion and dilution properties. These properties enlarge the results of absolute difference and highlighted as clear squirrel per figure 9 and 10. This situation can occur in natural settings where elements like moving water, wind-blown foliage, or other moving animals might alter the background dynamically.

Three consecutive image absolute pixel difference give better result compared to MRF, as it handled each pixel separately, and produce clear result by effectively highlighting structure of animal as per output images. However, it required static foreground and background as this method strongly depends on pixel change.

(a)

(b)

(c)



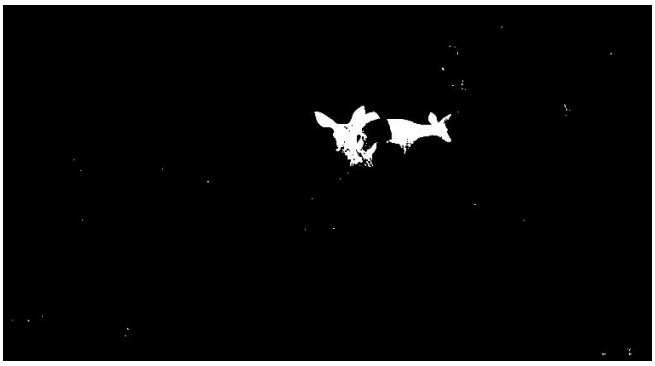
(d)

Figure 9 (a) First input image (b) Second input image (c) Third input image (d)Final output of Absolute Difference method

(a)

(b)

(c)



(d)

Figure 10 (a) First input image (b) Second input image (c) Third input image (d)Final output of Absolute Difference method

6.2.3 YOLOv8 results

The YOLOv8 neural network base object detection method is tested with 500 input images where 340 images are training, 100 testing and remaining are validation image dataset. By applying YOLOv8 on the given software and hardware set up, it took 3 hours to train the model.

In machine learning and neural network models the performance evaluation plays a critical role especially during model progression and to evaluate model accuracy, hence it is a necessity to implement quantifiable and precise performance metrics. This report analyzed the model accuracy by precision, recall and F1 score, evaluation matrixes used to measure classification model. A high precision value indicates a decreased false-positive rate. Precision is defined as the ratio of accurately detected positive classifications to the total anticipated positives. As a result, this statistic offers crucial information about how well our partial quadrant design categorizes the image content. The precision is defined as:

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

Recall examines a model's capacity to locate every relevant case in a dataset. It is the proportion of all observations made in the actual class to all positively predicted observations. A model with a high recall rate is one that catches a significant percentage of positive labels. Recall is useful in situations where capturing as many positive results as possible is essential, even if it means increasing the number of false positive mistakes (for an instance disease screening), the recall is defined as:

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

The weighted average (harmonic mean) of Precision and Recall is known as the F1 Score. This score accounts for both false positives and false negatives. It is very helpful in imbalanced classes. An F1 score of high indicates a strong equilibrium between Precision and Recall, rendering it a superior metric over accuracy, particularly when dealing with unbalanced datasets. F1 score is defined as:

$$F1\ Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

F1 score utilize for striking a balance between recall and precision. The F1 Score is applied when there is an unequal class distribution (many actual negatives), and you want to strike a

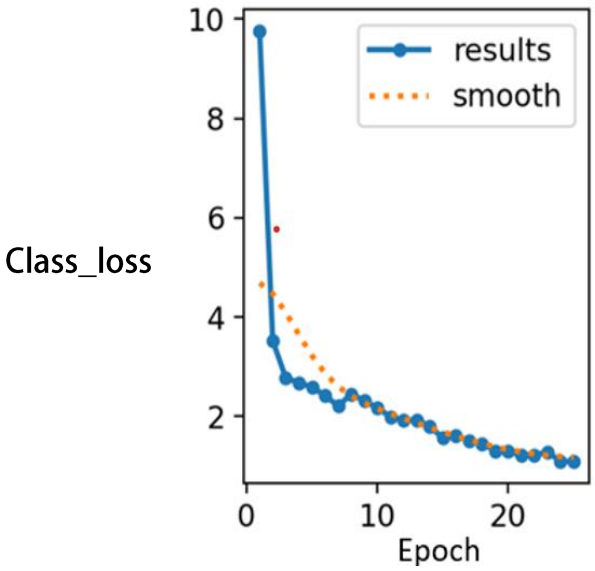


Figure 11 train/Class_loss

compromise between precision and recall. Below are training graphs analysis, which describes

how well model is trained with every epoch. The graph of train/Class_loss in figure 11 shows classification loss during each epoch of neural network, the class_loss shows how accurately model's prediction matched with class 0 or class 1. Actual classification loss mentioned by blue

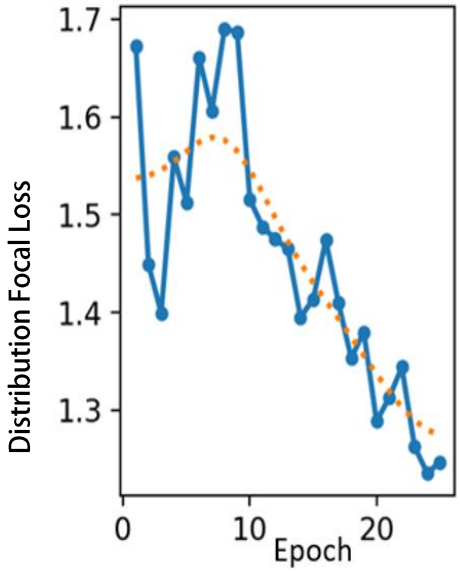


Figure 12 train/Distribution Focal Loss

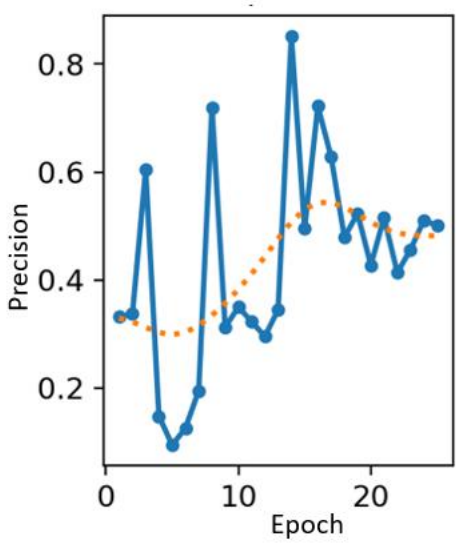


Figure 13 Matrix/Precision

line, it shows that model is quickly learning from training data where orange line shows overall classification loss by smoothing out fluctuations. The figure 12 train/Distribution Focal Loss will

treat the model state as a probability distribution problem and improve accuracy of bounding box prediction. During initial epochs it fluctuate but overall, it shows decreasing trend as it can extract detailed features, and overall models' ability to learn in enhancing. Figure 13 show precision

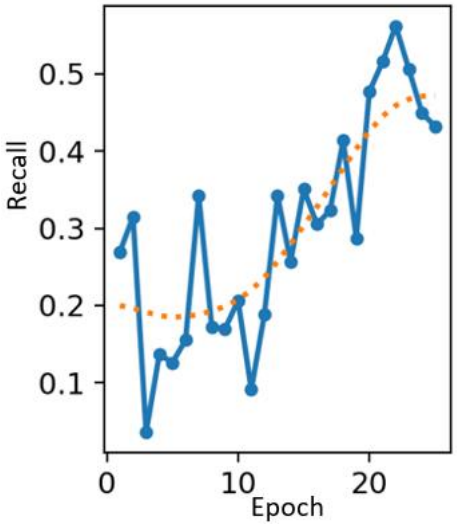


Figure 14 Matrix/Recall

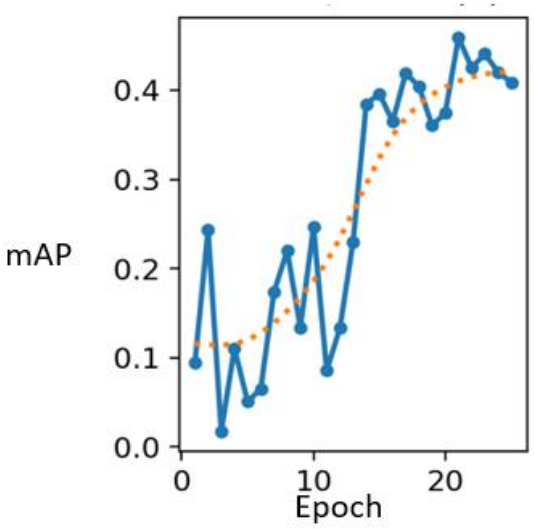


Figure 15 Matrix/mAP

matrix over epochs, Precision shows how often the model is correct when predicting the small bird. It is one of the useful measures when model is overfitting and find all objects of the target class. In Matrix/recall figure 14 shows an upward trend, which shows progressive improvement

in models' capability to detect all positives. The Mean average precision (matrix/mAP) provides an overall measure of model's accuracy across both classes. The figure 15 matrix/mAP shows model accuracy is increased to identify and localize object with epochs. Initial recall is high, which shows that it catches more true positives, however as threshold increases the recall decreases, shows that model struggles more in detecting actual bird.

It appears from the training graph analysis that the neural network model is successfully enhancing and adjusting over time to the complexities of object detection tasks. Both the distribution focal loss and the classification loss show that the model is getting more stable and accurate in predicting objects. The classification loss highlights the model's quick early learning, while the focal loss emphasizes how good the model is getting at predicting bounding boxes. The trends in precision and recall show that although the model overfits sometimes, it continuously improves in identifying all positives, which is essential for accurate object detection in a variety of situations.

The model's improved capacity to precisely identify and localize objects—achieving a balance between recognizing true positives and maintaining high precision—is confirmed by the increased trend in mean average precision (mAP) over epochs. To summarize, these observations show a positive direction for the model's evolution, indicating that high levels of accuracy and dependability could be attained in real-world applications with further training and modifications.

After training graph analysis below are sample output from YOLOv8, where in figure 16 and 17 a small bird is highlighted with bounding box with confidence 0.53. This confidence score is an essential parameter because it expresses how confident the model is in its predictions, which

is a critical aspect in determining whether detections are accepted as true positives or rejected as false positives. Higher confidence scores indicate more accurate and reliable detections.

The confidence score of 0.54 indicates an acceptable degree of certainty, which is important

Figure 16 YOLOv8 output image 1

Figure 17 YOLOv8 output image 2

for decision making, particularly in environmental monitoring where it is critical to separate real species from background disturbances. Reducing manual review workloads and enhancing automation dependability require striking a balance between recall (the model's capacity to identify all relevant instances) and precision (the model's accuracy in its detections), which can be achieved by adjusting the confidence threshold. We can achieve higher confidence thresholds

which produced more accurate predictions but lead to a reduction in the total number of birds spotted.

As a result, this confidence score influences not just how quickly detected events are

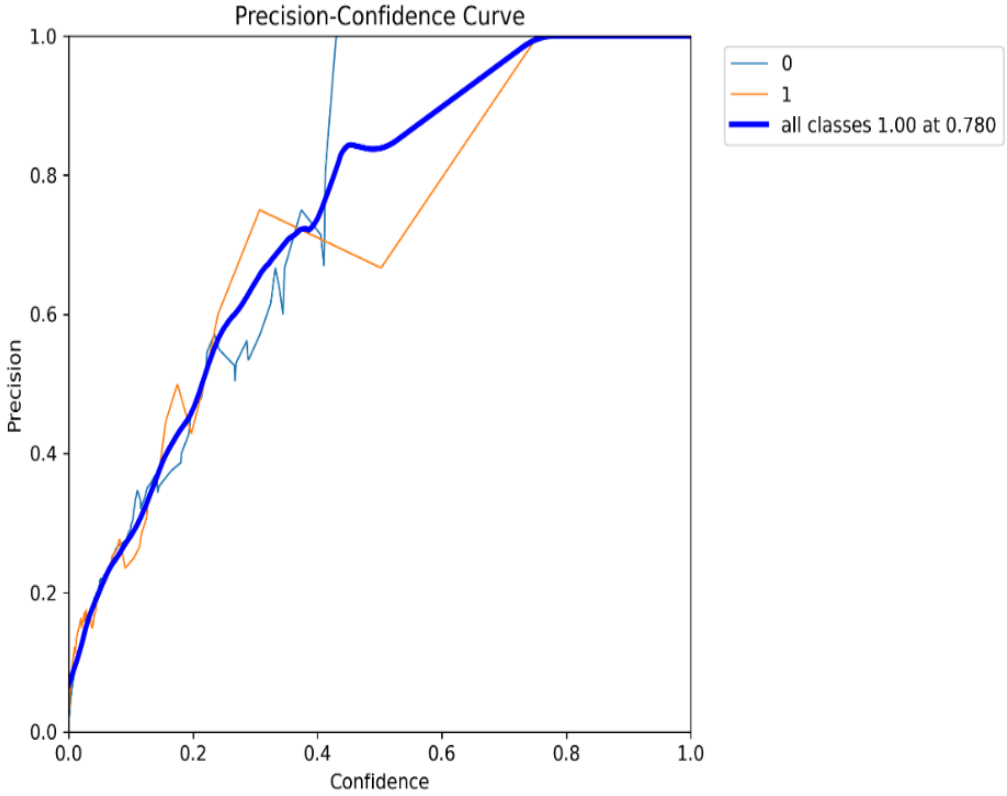


Figure 18 Precision-Confidence Curve

responded to, but also how well the model performs over time and how effectively resources are distributed depending on the certainty of detection. In figure: 18 Precision-Confidence Curve, precision increases as confidence is increases, which indicates that model have fewer false positives at higher thresholds. Initially lower precision for class 1, shows that model is prone to false positives on bird prediction, due to similar features in the backgrounds.

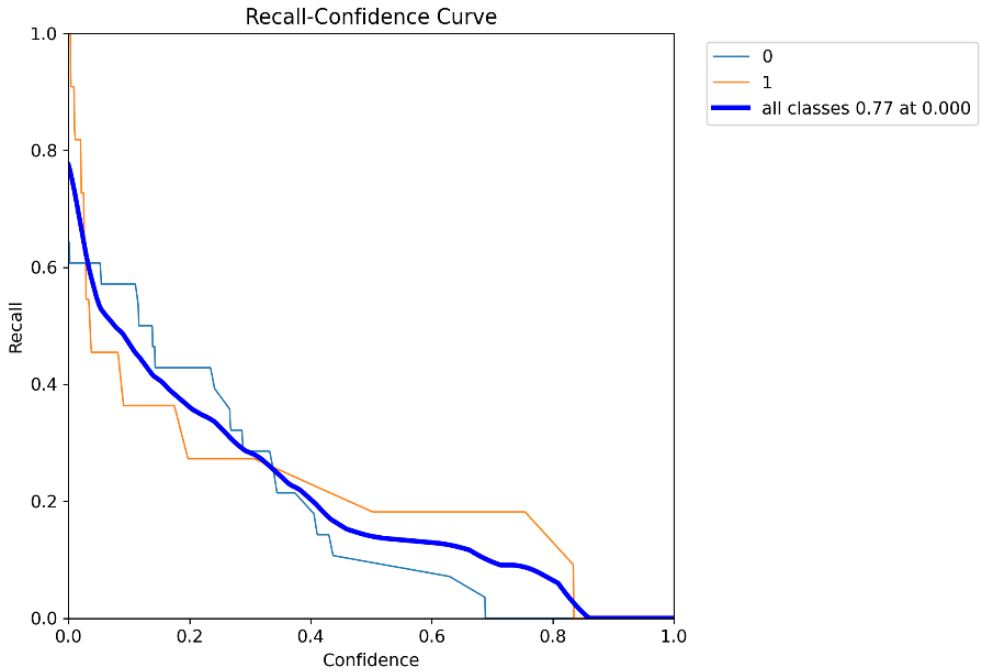


Figure 19 Recall-Confidence Curve

Recall is particularly important in applications where an actual instance (like not spotting a bird when it is present) carries a significant penalty. As confidence levels rise in Figure 19 Recall-Confidence Curve, the graph displays a usual downward trend in recall because stronger thresholds tend to reject more detections, including some right prediction.

The all-classes curve's peak recall at the lowest confidence level suggests that lowering the threshold will increase the number of false positives while optimizing the model's capacity to identify all birds. F1 score in figure 20 is harmonic mean of precision and recall, for class-1 the curve initially increases and then drops as confidence increases, shows that many true positive for class-1 are identified with average confidence. However, for class-0 model is more confident to detect non-bird objects. Overall, the model demonstrates a competent ability to identify the presence of birds (class 1). The optimal performance, as indicated by the peak points in the precision and F1 score graphs, occurs at moderate to high confidence thresholds.

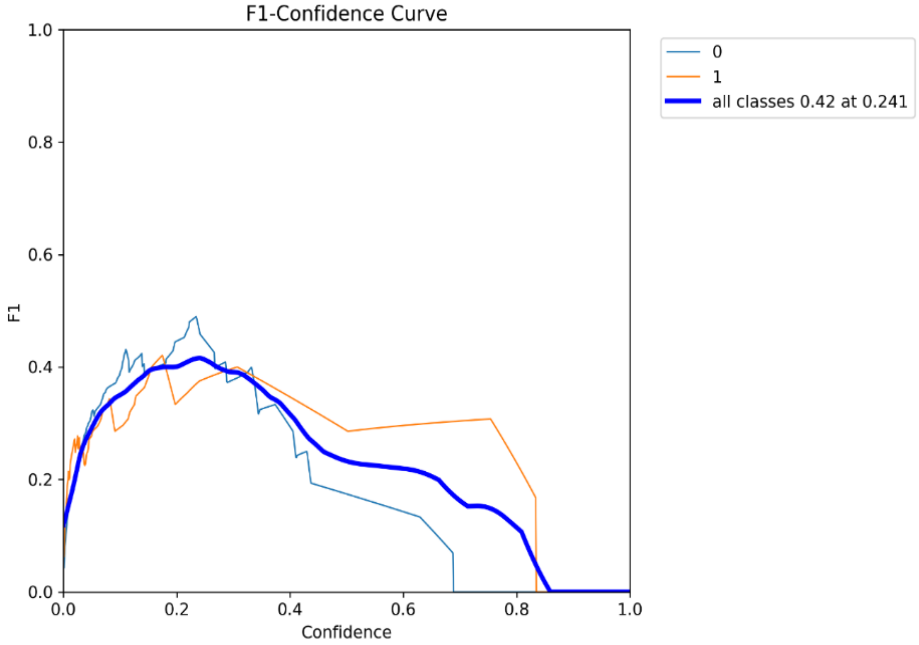


Figure 20 F1-Confidence Curve

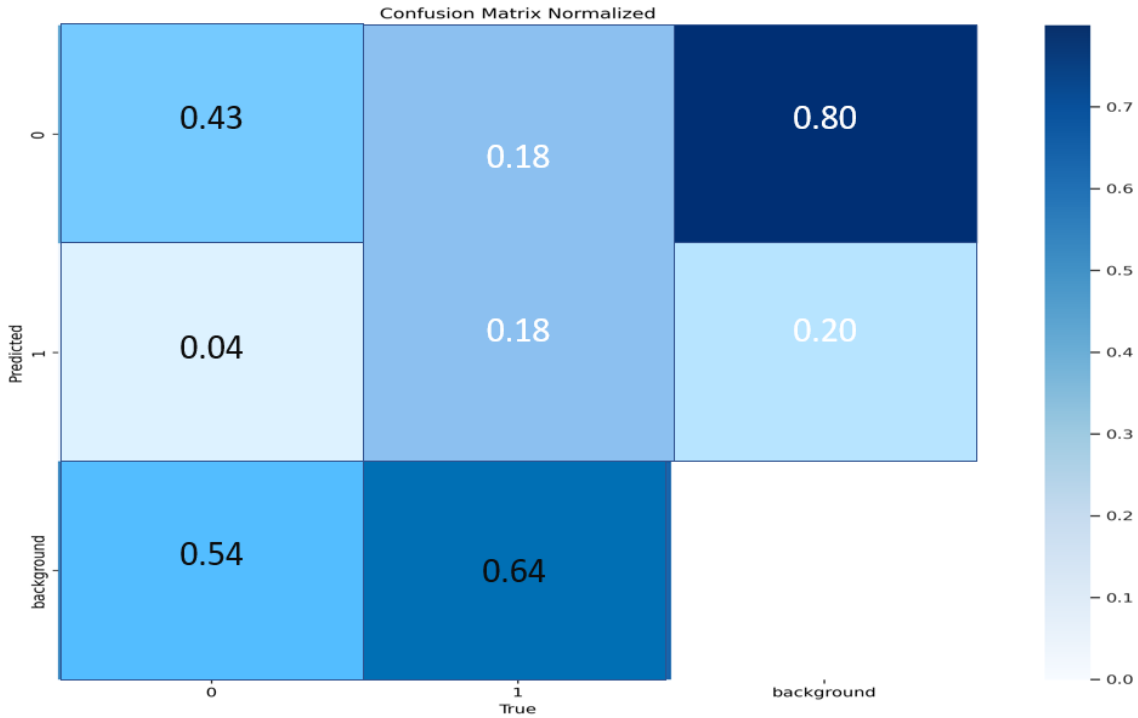


Figure 21 Confusion Matrix

A confusion matrix describes the performance of the classification model on a set of test images for which true values are unknown. It gives a visualization of the test image performance and gives a numerical analysis of how well the model is predicting actual true result, how many predictions misclassified. This matrix visualization is a guide to further improvement in research. Here, in figure 21 rows represent predicted class by model and columns represent True (actual) class, where class 0 means non-bird object and class 1 means bird object. Background is a baseline or default which does not belong to any class. Here true positive is 0.43 where the model correctly predicted class 0 when the truth was class 0, false positive is 0.18 model incorrectly predicted class 1 when the truth was class 0, false negative is 0.04 where the model incorrectly predicted class 0 when the truth was class 1 and true positives is 0.18 where the model correctly predicted class 1 when the truth was class 1.

The model easily predicts non bird objects compared to small bird; this indicates bias towards predicting class 0. The "Background" values (0.54 and 0.64) indicate that model classify ambiguous cases as background rather than making FP error on class 1. To conclude the whole analysis model identifies small birds with TP 0.43 which shows the capability of detailed feature extraction of YOLOv8 compared to the absolute difference between three images which can only identify large animals and struggles in detecting small pixel cluster of a bird. This discussion led to the conclusion of the resort in the next chapter.

CHAPTER 7: CONCLUSION

In conclusion, the research report presents an in-depth study of two different methodologies for small bird detection: non neural network-based Markov Random Field (MRF) technique and neural network based YOLOv8 algorithm. The objective of the research was to assess the efficiency of these approaches to identify small birds in complex and challenging background of an image.

The MRF method is beneficial to model spatial dependencies and contextual relationship between pixels of an image. Although MRF is effective in local pixel comparison based on previous knowledge, it struggled with detecting small objects due to limited features extraction. The proposed MRF method produced reliable image segmentation but was less effective in distinguishing the small birds from background.

The absolute pixel difference method processes three consecutive images with Gaussian blur, histogram equalization and morphological filtering. This method proved effective in showcasing animal structure in static environment. As it relied more on solid background and foreground it struggled in dynamic settings.

The YOLOv8 algorithm, a neural network-based approach showed superior performance in small bird detection compared to MRF. It demonstrated higher confidence and precision in identifying birds, easily managed complex patterns and pixel clusters that posed challenges for traditional algorithms. The better feature extraction and detection highlights the power of neural networks.

YOLOv8 outperformed the Markov Random Field (MRF) technique to detect small birds, due to the depth of network in feature extraction. YOLOv8 demonstrated a higher confidence in detecting birds, setting a threshold confidence of over 0.35, effectively identifying smaller objects which MRF struggled with. The neural network's ability to process and learn from complex image data allowed YOLOv8 to excel where non-neural network methods like MRF could not. This highlights the power of deep neural networks. YOLOv8's advanced feature extraction capabilities outperformed traditional methods, making it a powerful tool for detecting small and distant birds in complex backgrounds.

7.1 Future Work

YOLOv8 tests images with decent confidence, however its performance can be improved by training on more datasets. YOLOv8's performance depends on the quality and diversity of the training dataset, hence with more diverse data it will train efficiently. On the other hand, improving the MRF technique by introducing inference methods like LayoutCRF has the potential to yield better results in small object detection. By addressing these areas, the detection of small and distant birds in complex environments can be significantly improved, contributing to more effective wildlife monitoring and conservation efforts.

References:

(Huang, 2018) R. Huang, J. Pedoeem and C. Chen, "YOLO-LITE: A Real-Time Object Detection Algorithm Optimized for Non-GPU Computers," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 2503-2510, doi10.1109/BigData.2018.8621865

(Schneider, 2018) S. Schneider, G. W. Taylor, and S. Kremer, "Deep Learning Object Detection Methods for Ecological Camera Trap Data," 2018 15th Conference on Computer and Robot Vision (CRV), Toronto, ON, Canada, 2018, pp. 321-328, doi: 10.1109/CRV.2018.00052

(Liu, 2018) C. Liu, Y. Tao, J. Liang, K. Li and Y. Chen, "Object Detection Based on YOLO Network," 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 2018, pp. 799-803, doi: 10.1109/ITOEC.2018.8740604

(Nguyen, 2019) N. -D. Nguyen, D. -H. Bui, and X. -T. Tran, "A Novel Hardware Architecture for Human Detection using HOGSVM Co-Optimization," 2019 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), Bangkok, Thailand, 2019, pp. 33-36, doi: 10.1109/APCCAS47518.2019.8953123

(Wang, 2009) X. Wang, C. Jingdong, X. Huosheng and C. Xi, "Adaptive method for infrared small target detection based on gray-scale morphology and backward cumulative

histogram analysis," 2009 International Conference on Information and Automation, Zhuhai/Macau, China 2009, pp. 173-177, doi: 10.1109/ICINFA.2009.5204915

(Karanth, 1995) K. U. Karanth, “Estimating tiger Panthera tigris populations from camera-trap data using capture recapture models,” Biological conservation, vol. 71, no. 3, pp. 333–338, 1995, doi: 10.1016/0006-3207

(Gysel, 1956) L. W. Gysel and E. M. Davis, “A simple automatic photographic unit for wildlife research,” The Journal of Wildlife Management, vol. 20, no. 4, pp. 451–453, 1956, doi: 10.2307/3797161

(Li, 2021) S. Li, Y. Li, Y. Li, M. Li and X. Xu, "YOLO-FIRI: Improved YOLOv5 for Infrared Image Object Detection," IEEE Access, vol. 9, pp. 141861-141875, 2021, doi: 10.1109/ACCESS.2021.3120870

(Christin, 2021) S. Christin, E. Hervet, N. Lecomte and H. Ye, "Going further with model verification and deep learning," Methods Ecol. Evol., vol. 12, no. 1, pp. 130–134, 2021, doi: 10.1111/2041-210X.13494

(Godley, 2008) B. Godley, J. Blumenthal, A. Broderick, M. Coyne, M. Godfrey, L. Hawkes, and M. Witt, “Satellite tracking of sea turtles: Where have we been and where do we go next?” Endangered Species Research, vol. 4, pp. 3–22, 2008, doi: 10.3354/esr00060

(Girshick, 2014) R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 580-587, doi: 10.1109/CVPR.2014.81

(Girshick, 2015) R. Girshick, "Fast R-CNN," in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169

(Beery, 2019) S. Beery, D. Morris, and S. Yang, "Efficient pipeline for camera trap image review," ArXiv:1907.06772 (Cs), 2019, doi: 10.48550/arXiv.1907.06772

(Beery S. , 2020) S. Beery, G. Wu, V. Rathod, R. Votel and J. Huang, "Context R-CNN: Long Term Temporal Context for Per-Camera Object Detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 13072-13082, doi: 10.1109/CVPR42600.2020.01309.

(Hulbert, 2001) I. A. Hulbert and J. French, “The accuracy of GPS for wildlife telemetry and habitat mapping,” Journal of Applied Ecology, vol. 38, no. 4, pp. 869–878, 2001, doi: 10.1046/j.1365-2664.2001.00624.x

(Yan, 2023) X. Yan, B. Shen, and H. Li, "Small Objects Detection Method for UAVs (Unmanned Aerial Vehicles) Aerial Image Based on YOLOv5s," 2023 IEEE 6th International

Conference on Electronic Information and Communication Technology (ICEICT), Qingdao, China, 2023, pp. 61-66, doi: 10.1109/ICEICT57916.2023.10245156

(Ultralytics, 2023) Home - Ultralytics YOLOv8 Docs

(Du, 2018) P. Du, X. Qu, T. Wei, C. Peng, X. Zhong and C. Chen, "Research on Small Size Object Detection in Complex Background," *2018 Chinese Automation Congress (CAC)*, Xi'an, China, 2018, pp. 4216-4220, doi: 10.1109/CAC.2018.8623078

(Kim, 2022) H. M. Kim, J. H. Kim, K. R. Park, and Y. S. Moon, "Small Object Detection using Prediction head and Attention," *2022 International Conference on Electronics, Information, and Communication (ICEIC)*, Jeju, Korea, Republic of, 2022, pp. 1-4, doi: 10.1109/ICEIC54506.2022.9748393.

(Li, 2020) W. Li, X. Tan and Z. Wang, "Small Object Detection of Table Tennis Based on Deep Learning Network," 2020 International Conference on Computer Science and Management Technology (ICCSMT), Shanghai, China, 2020, pp. 149-152, doi: 10.1109/ICCSMT51754.2020.00036.

(Redmon, 2016) J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

(Blake, 2011) A. Blake, P. Kohli and C Rother, (Eds.) (2011). Markov Random Fields for Vision and Image Processing. The MIT Press, 2011, ISBN: 978-0-262-29835-3, doi: 10.7551/mitpress/8579.001.0001

J. L. Marroquin, E. Arce and S. Botello, "Markov random measure fields for image analysis," ICIP 2002 International Conference on Image Processing, Rochester, NY, USA, 2002, pp. I-765-I-768, doi: 10.1109/ICIP.2002.1038137

(Ait-Aoudia, 2011) S. Ait-Aoudia, R. Mahiou and E. Guerrout, "Evaluation of Volumetric Medical Images Segmentation Using Hidden Markov Random Field Model," 2011 15th International Conference on Information Visualisation, London, UK, 2011, pp. 513-518, doi: 10.1109/IV.2011.83

(Geman, 1984) S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-6, no. 6, pp. 721-741, 1984, doi: 10.1109/TPAMI.1984.4767596

(Cross, 1983) G. R. Cross and A. K. Jain, "Markov Random Field Texture Models," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-5, no. 1, pp. 25-39, 1983, doi: 10.1109/TPAMI.1983.4767341

(Lee, 2017) J. Lee, J. Bang and S. -I. Yang, "Object detection with sliding window in images including multiple similar objects," 2017 International Conference on Information and

Communication Technology Convergence (ICTC), Jeju, Korea (South), 2017, pp. 803-806, doi: 10.1109/ICTC.2017.8190786

(Wang, 2009) X. Wang, C. Jingdong, X. Huosheng and C. Xi, "Adaptive method for infrared small target detection based on gray-scale morphology and backward cumulative histogram analysis," 2009 International Conference on Information and Automation, Zhuhai/Macau, China, 2009, pp. 173-177, doi: 10.1109/ICINFA.2009.5204915

(Mahum, 2022) D. -e. -M. Nisar, R. Mahum, T. Azim and N. -u. -h. Shah, "Proteins Classification Using An Improve Darknet-53 Deep Learning Model," 2022 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), Karachi, Pakistan, 2022, pp. 1-6, doi: 10.1109/MAJICC56935.2022.9994209

(Mishra, 2014) S. Mishra, R. S. Sengar, R. K. Puri and D. N. Badodkar, "Efficient motion blur parameters estimation under noisy conditions," 2014 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, India, 2014, pp. 1-5, doi: 10.1109/ICCIC.2014.7238308

(Bethel, 2017) G. N. B. Bethel, T. V. Rajinikanth and S. V. Raju, "An Improved Analysis of Heart MRI Images using the Morphological Operations," 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore, India, 2017, pp. 891-897, doi: 10.1109/CTCEEC.2017.8455050

(Chauhan, 2011) R. Chauhan and S. S. Bhadoria, "An Improved Image Contrast Enhancement Based on Histogram Equalization and Brightness Preserving Weight Clustering Histogram Equalization," 2011 International Conference on Communication Systems and Network Technologies, Katra, India, 2011, pp. 597-600, doi: 10.1109/CSNT.2011.128

(Yu, 2006) Shun-Zheng Yu and H. Kobayashi, "Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden Markov model," in IEEE Transactions on Signal Processing, vol. 54, no. 5, pp. 1947-1951, 2006, doi: 10.1109/TSP.2006.872540

(Ardo, 2007) H. Ardo, K. Astrom and R. Berthilsson, "Real Time Viterbi Optimization of Hidden Markov Models for Multi Target Tracking," 2007 IEEE Workshop on Motion and Video Computing (WMVC'07), Austin, TX, USA, 2007, pp. 2-2, doi: 10.1109/WMVC.2007.33