Aristotle University of Thessaloniki

Faculty of Sciences

School of Informatics

Bsc

Detection of canine and feline intestinal parasites using computer vision

Georgios Gerakis

ID: 3297

Supervisor: Professor Grigorios Tsoumakas

June 30, 2025

Thanks To

I am deeply grateful to my supervisor Tsoumakas Grigorios whose guidance has been invaluable throughout my research. My sincere thanks also go to Dimitriadis Dimitrios for his guidance, support and always being available to offer guidance when needed. I am also thankful to Constantina Tsokana for the contribution of her valuable expertise and collection of data.

Lastly, I would like to express my gratitude to my family, without whose support this work would have been impossible.

Abstract

The early and accurate identification of intestinal parasites in companion animals is critical for both veterinary care and public health. Traditional diagnostic methods, while effective, are labor-intensive, time-consuming, and require expert interpretation, which limits their accessibility and scalability, especially in point-of-care contexts. In this thesis, we propose an automated pipeline leveraging deep learning and computer vision techniques to detect and classify common canine and feline intestinal parasites directly from microscopy images. An object detection model was fine-tuned on a curated, annotated dataset of fecal smear images acquired under varied real-world conditions. The system demonstrated high performance across several evaluation metrics, indicating strong generalization to unseen samples and robustness in complex visual environments. Key contributions of this work include dataset construction, data augmentation tailored to microscopic features, and real-time inference capability. The results suggest significant promise for deploying AI-powered tools in veterinary diagnostics, potentially enabling faster, more accessible, and cost-effective screening of zoonotic parasites in clinical and field settings.

Keywords: Intestinal Parasites, Microscopy Images, Computer Vision, Deep Learning, Object Detection

Περίληψη

Η έγκαιρη και ακριβής αναγνώριση εντερικών παρασίτων σε ζώα συντροφιάς είναι κρίσιμη τόσο για την κτηνιατρική φροντίδα όσο και για τη δημόσια υγεία. Οι παραδοσιακές διαγνωστικές μέθοδοι, παρότι αποτελεσματικές, είναι χρονοβόρες, απαιτούν εντατική εργασία και εξειδικευμένη ερμηνεία, γεγονός που περιορίζει την προσβασιμότητα και τη δυνατότητα κλιμάκωσής τους, ειδικά σε σημεία παροχής άμεσης φροντίδας. Στην παρούσα εργασία προτείνεται μια αυτοματοποιημένη διαδικασία που αξιοποιεί τεχνικές βαθιάς μάθησης και υπολογιστικής όρασης για την ανίχνευση και ταξινόμηση κοινών εντερικών παρασίτων σκύλων και γατών απευθείας από εικόνες μικροσκοπίου. Ένα μοντέλο εντοπισμού αντικειμένων προσαρμόστηκε σε ένα επιμελώς επισημασμένο σύνολο δεδομένων με εικόνες από επιχρίσματα κοπράνων, οι οποίες αποκτήθηκαν υπό ποικίλες συνθήκες του πραγματικού κόσμου. Το σύστημα επέδειξε υψηλές επιδόσεις σε διάφορους δείκτες αξιολόγησης, υποδεικνύοντας ισχυρή γενίκευση σε μη γνωστά δείγματα και ανθεκτικότητα σε περίπλοκα οπτικά περιβάλλοντα. Βασικές συνεισφορές της παρούσας μελέτης περιλαμβάνουν τη δημιουργία του συνόλου δεδομένων, τεχνικές εμπλοτισμού των δεδομένων προσαρμοσμένες στα μικροσκοπικά οπτικά γνωρίσματα, καθώς και τη δυνατότητα πρόβλεψης σε πραγματικό χρόνο. Τα αποτελέσματα υποδηλώνουν σημαντικές προοπτικές για την εφαρμογή εργαλείων τεχνητής νοημοσύνης στη διαγνωστική κτηνιατρική, με πιθανή επίτευξη ταχύτερης, πιο προσβάσιμης και οικονομικότερης ανίχνευσης ζωονοσογόνων παρασίτων σε κλινικά πεδία εφαρμογής.

Λέξεις κλειδιά: Εντερικά Παράσιτα, Εικόνες Μικροσκοπίας, Όραση Υπολογιστών, Βαθεία Μάθηση, Εντοπισμός Αντικειμένων

Contents

List of Figures

List of Tables

1 Introduction

1.1 Global Health and Companion Animals

Pet dogs and cats are often regarded as loyal companions, contributing positively to human emotional development, socialization, and overall well-being. However, they can also serve as reservoirs for various zoonotic parasites. Their role in transmitting infections to humans has been widely recognized on a global scale [1, 2, 3].

The distribution of canine and feline parasites in Europe has expanded significantly, largely due to human activities such as pet trading and animal protection efforts, which facilitate the movement of dogs across the continent. As human travel increases alongside population growth and rising financial status, many people choose to bring their companion animals, especially dogs, with them. This increased mobility contributes to the global spread of zoonotic parasites, as pets can carry and transmit infectious agents across different regions. The movement of dogs, whether for leisure, relocation, or trade, plays a crucial role in the dispersal of parasitic diseases. Ecological changes have also played a role, with increasing fox populations and the spread of the invasive raccoon dog, both of which serve as reservoirs for canine helminths. Additionally, large populations of free-roaming stray cats and dogs, especially in Southern Europe, contribute to a continuous infection risk for domestic pets [3, 4].

In China, rapid socioeconomic development and rising living standards have led to an increase in the number of dogs and cats kept as family pets. However, many of these animals, especially in urban and rural settings, roam freely, bringing them into close contact with humans. This close interaction poses a significant risk of zoonotic disease transmission, as stray and free-roaming pets can act as reservoirs for various parasites and infectious agents [5].

Intestinal parasitic infections are not only a major public health concern but also a significant socioeconomic burden. These infections can lead to malabsorption, diarrhea, hemorrhage, impaired work capacity, reduced growth rates, and cognitive deficits, particularly in children. The resulting health complications can diminish productivity, increase healthcare costs, and perpetuate cycles of poverty, especially in resource-limited communities. Advancements in diagnostic technologies are crucial to mitigating their impact on public health and economic development [6].

Helminth larvae in dogs reside in the intestine, where they produce thousands of eggs that are excreted in feces, leading to environmental contamination. Transmission occurs through contaminated water, ingestion of improperly washed or cooked vegetables, and hand-to-mouth contact, particularly in children playing in contaminated soil. As a result, helminth eggs, protozoan cysts, and oocysts excreted in the feces of infected animals become a primary source of infection for both animals and humans [6, 7].

Some of the most prevalent endoparasites in dogs and cats include several species of *Giardia*, *Toxocara*, *Ancylostoma* (Hookworm) and *Cystoisospora*. With the exception of Cystoisospora, all of the aformentioned parasite species are of zoonotic nature and thus are a public health concern as they can be transmitted to humans [8, 9, 10, 11, 4, 12, 13].

Human toxocariasis is a zoonotic infection caused by the roundworms *Toxocara canis* (from dogs) and *T. cati* (from cats). Transmission occurs when humans accidentally ingest embryonated eggs through contaminated soil, food, surfaces, or direct contact with infected animals. The infection can lead to various clinical syndromes, including visceral larva migrans, ocular larva migrans, eosinophilic meningoencephalitis, covert toxocariasis, and neurotoxocarosis [14, 15, 16]. Toxocariasis is among the most widespread human parasitic infections globally, with particularly high prevalence reported in developing countries. The extensive presence of *Toxocara* species in dogs and cats contributes to environmental contamination with *Toxocara* eggs, posing a significant public health threat.

Ancylostoma species are recognized as endemic and widely distributed hookworms affecting dogs and cats throughout Asia. Molecular studies in the region have identified *Ancylostoma ceylanicum* as one of the most prevalent hookworm species infecting humans. Although natural human infections have been documented in nearly all areas where these hookworms are endemic in animals, most of these reports lack accompanying clinical data [17]. Larval infections transmitted from dogs and cats to humans can result in cutaneous larva migrans, a condition commonly associated with hookworms. In less frequent cases, such infections may also cause eosinophilic pneumonitis, eosinophilic enteritis, localized myositis, folliculitis, erythema multiforme, or even ophthalmological complications [18].

Giardiasis is a significant intestinal protozoan infection that poses public health challenges in many developing countries, as well as in some developed nations. *Giardia lamblia* is a known zoonotic pathogen responsible for giardiasis in humans and various mammals, including dogs and cats. The illness often presents with symptoms such as watery diarrhea, vomiting, abdominal discomfort, malabsorption, and other related issues, especially in young children. Epidemiological studies indicate that *G. lamblia* is a leading cause of parasitic diarrhea in children, particularly in regions where drinking water and fresh vegetables are contaminated with sewage, and where food is commonly purchased from street vendors. The potential for zoonotic transmission of *G. duodenalis* from dogs and cats to humans is notable, given the frequent presence of zoonotic assemblages and the close interactions between pets and people [19, 20].

Coccidiosis is a gastrointestinal infection caused by *Cystoisospora* species. It is commonly found in dogs and cats worldwide, with puppies and kittens being especially susceptible. Infection typically occurs through ingestion of sporulated oocysts from the environment or direct contact with contaminated feces. While exposure to *Cystoisospora* is widespread, it often does not result in noticeable illness. However, in young animals, clinical symptoms such as diarrhea, dehydration, and weight loss can appear even before oocysts are detectable in feces. Diagnosis is made by examining fecal samples for the presence of coccidian oocysts. Importantly, the Cystoisospora species that infect dogs and cats are not considered zoonotic and do not pose a risk to humans [21, 22].

Manual microscopy fecal testing still remains the primary method for identifying intestinal parasites [23, 24]. Veterinarians recommend at least an annual screening to safeguard both pet and owner health [25, 18]. However, traditional fecal examination methods are labor-intensive and time-consuming, making them impractical for large-scale testing. Manual analysis requires extensive effort to extract

relevant features, interpret findings, and diagnose the specific parasite eggs present in the sample [24, 26].

To address these challenges, an automated method for detecting and identifying intestinal parasites in canine fecal samples is essential. Such a method would significantly accelerate the diagnostic process, improving efficiency and accuracy in fecal analysis.

1.2 Computer Vision in Medical Imaging

Over the past few years, computer vision has transitioned from rule-based image processing systems to powerful, data-driven models capable of near-human-level understanding of visual scenes. This shift has been driven primarily by the advent of deep learning, particularly convolutional neural networks (CNNs), which have dramatically improved performance in tasks such as object detection, classification and segmentation.

Computer vision has shown transformative potential in medical imaging, including radiology [27], histopathology [28], dermatology [29], and ophthalmology [30, 31]. These domains benefit from the automation of visual tasks that are typically time-consuming, subjective, and prone to human error. Models trained on large annotated datasets have demonstrated capabilities in detecting pathologies, quantifying lesions, and segmenting anatomical structures with a high degree of accuracy. In these applications, computer vision not only accelerates diagnosis but also aids in decision support, standardization, and remote screening [32, 33].

The integration of computer vision into parasitology has significantly improved the detection and classification of intestinal parasites. These methods not only enhance diagnostic accuracy but also increase efficiency, making large-scale screening more feasible. Studies have demonstrated that CNN-based models can closely match the results of traditional methods in detecting intestinal parasites in stool specimens, paving the way for more reliable and rapid diagnostic tools in clinical and laboratory settings [34, 35, 36, 37, 38].

Developments in lightweight architectures such as the YOLO (You Only Look Once) [39] family of models, enable deployment on resource-constrained platforms, making them suitable for use in low-infrastructure settings such as rural clinics or field laboratories. The deployment of real-time object detectors can fill a critical gap: automating detection and classification of parasite species in raw or minimally processed samples with high throughput and minimal delay [40].

This work aims to fill the literature gap in veterinary parasitology by contributing a modern object detection solution to the task of intestinal parasite detection. It addresses the challenges of real-world variability, class imbalance, and the need for explainability in clinical contexts. To this end, a novel dataset to train and test deep learning models is presented, from samples provided by the Laboratory of Parasitology and Parasitic Diseases, School of Veterinary Medicine, Faculty of Health Sciences, Aristotle University of Thessaloniki.

1.3 Thesis Structure

This thesis is structured to address both the theoretical underpinnings and practical implementations of computer vision-based approaches for detecting intestinal parasites in canine and feline microscopy images. The chapters are arranged as follows:

Chapter 2 introduces the foundational background necessary for understanding the core components of the research. It discusses essential topics such as deep learning, computer vision and the architecture of YOLO models. These principles provide the theoretical basis for the advanced image analysis and classification techniques developed and employed in later stages of this study.

Chapter 3 offers an in-depth survey of current developments in computer vision, with a particular focus on applications in biomedical and veterinary diagnostics. Special attention is given to the use of object detection methods within parasitology, underscoring the potential of automated microscopy analysis in real-time diagnostic workflows. The chapter situates the research within the broader context of vision-based pathogen identification and explores how modern detection pipelines can be adapted for veterinary use cases.

Chapter 4 touches on the motivation of this study and gives a clear definition and formulation for the problem at hand. In addition, the dataset utilized in this work is presented, detailing the procedures followed for data acquisition, preprocessing, annotation, and quality control. Finally, it outlines the methodological framework adopted in the study elaborating on the rationale behind selecting specific algorithms suited to the domain of parasite localization and classification in microscopy images.

Chapter 6 presents the practical implementation of the proposed detection pipeline. It includes a detailed explanation of the experimental setup, including hardware, hyperparameter settings, data augmentation strategies, and data splits. The chapter also discusses the performance results across multiple metrics, such as precision, recall, mean average precision (mAP). In addition, it addresses model interpretability through a saliency-based visualization tool, ensuring transparency and accountability in a medical and diagnostic setting.

Chapter 7 concludes the thesis by synthesizing the primary findings and reflecting on their implications for veterinary parasitology and public health. It highlights the potential of computer vision systems to provide scalable, accurate, and rapid diagnostic support in resource-limited clinical environments. The chapter also discusses the current limitations encountered in the study and offers recommendations for future research, including directions for improving model generalization and real world deployment.

2 Background

In recent years, deep learning has emerged as a transformative approach for solving complex problems in computer vision, enabling machines to interpret and analyze visual information with unprecedented accuracy. At the core of this revolution are CNN, which are particularly effective at learning spatial hierarchies of features from image data. CNNs have driven major breakthroughs in tasks such as image classification, object detection, and segmentation. Building upon these advances, specialized architectures like YOLO have been developed to achieve real-time object detection by unifying localization and classification into a single, efficient framework. This chapter introduces the foundational concepts of deep learning and CNNs, presents the domain of computer vision, its goals and applications, followed by an in-depth review of the most recent YOLO architecture.

2.1 Deep Learning

Deep learning, a specialized subset of machine learning, utilizes artificial neural networks, inspired by the way the human brain processes information. They consist of multiple layers, each interpreting the input data at different levels of abstraction.

Conventional machine learning techniques typically require multiple sequential steps, including pre-processing, feature extraction, careful feature selection, learning, and classification. Among these, feature selection plays a crucial role, as biased selection can lead to incorrect class distinctions. In contrast, deep learning automates the feature learning process, mapping inputs directly to specific labels, eliminating the need for manual feature engineering.

The rise of deep learning in recent years has been driven by the explosion of big data, enabling remarkable advancements across various machine learning tasks. Its continuous evolution has simplified and enhanced fields such as image super-resolution, object detection, and image recognition. Notably, deep learning has surpassed human performance in tasks like image classification, reinforcing its transformative impact on artificial intelligence and data-driven decision-making.

In medical diagnostics, deep learning techniques process vast amounts of medical imaging data, enabling the classification of images into specific categories based on learned patterns. These models enhance accuracy and efficiency by automatically extracting relevant features, reducing reliance on manual analysis, and improving diagnostic consistency across diverse medical datasets.

A deep-learning architecture consists of multiple layers of simple, trainable modules, most of which perform nonlinear transformations on their inputs. In general, a neural network consists of multiple processing units organized into three main layers: the input layer, hidden layers, and output layer. With varying depths which can potentially reach in the thousands, these architectures can model highly complex input-output relationships. This allows them to be highly sensitive to subtle details while remaining robust to variations in background, pose, lighting, and surrounding objects.

Multilayer architectures can be effectively trained using stochastic gradient descent, with gradients computed through the backpropagation algorithm. Backpropagation applies the chain rule for deriva-

tives, propagating gradients backward from the output layer to the input layer. This step-by-step approach enables efficient weight updates for each layer.

A particular type of neural network that is both easier to train and more effective at generalizing than fully connected networks is the CNN. CNNs have demonstrated remarkable success in image recognition tasks and have become a standard approach in the field of computer vision. Their specialized architecture, which leverages spatial hierarchies and shared parameters, makes them particularly well-suited for analyzing visual data.

2.2 Computer Vision

Computer Vision is an interdisciplinary domain focused on enabling machines to perceive, interpret, and make decisions based on visual input from the world. By harnessing advancements in artificial intelligence and machine learning, computer vision seeks to replicate and automate tasks traditionally performed by human vision. This subsection outlines essential concepts and common tasks in the field.

Computer vision integrates methods from image processing, machine learning, and deep learning to analyze images and video data. Its primary goal is to extract valuable information and insights from visual content. Core tasks within computer vision include image classification, object detection, pose estimation, and image segmentation.

2.2.1 Computer Vision Goals and Applications

Image classification involves assigning a single categorical label to an entire image based on its visual content. Typically, the image is represented as a tensor $x \in \mathbb{R}^{H \times W \times C}$, where H is the height, W is the width, and C is the number of color channels (such as 3 for RGB). The objective is to predict a class label $\hat{y} \in \{1, 2, \ldots, K\}$, where K denotes the total number of distinct categories in the dataset.

The fundamental formulation of the classification task is:

$$\hat{y} = f(x), \quad f : \mathbb{R}^{H \times W \times C} \to \{1, 2, \ldots, K\}$$

The objective is to learn a function $f(x)$ that maps an input image x to one of the predefined class labels. This function is trained through supervised learning, where each training image comes with a known label, allowing the model to learn from correct examples.

A major challenge in this task is learning discriminative features from the image that help in correctly identifying its class. These features can include patterns like shapes, edges, textures, or more complex visual structures. Deep learning models, such as CNNs and Visual Transformers, are capable of automatically learning such features through their layered architecture.

To guide the training process, a loss function is used—most commonly, the cross-entropy loss, which measures the difference between the predicted probability distribution over classes and the actual label. The model adjusts its parameters to minimize this loss, thereby improving its classification accuracy on both training and unseen data.

Object detection expands on image classification by not only identifying what objects are in an image, but also where they are located. For each detected object, the model outputs a bounding box that tightly encloses the object, typically defined by the coordinates of the top-left and bottom-right corners $(x_i^{\text{tl}}, y_i^{\text{tl}})$ and $(x_i^{\text{br}}, y_i^{\text{br}})$, respectively. For each bounding box, the model also outputs its corresponding class label.

The goal is to learn a function

$$f(x) = (\hat{c}_i, \hat{b}_i)$$

where $\hat{c}_i$ is the predicted class for the i-th object, and $\hat{b}_i$ is its predicted bounding box.

This dual task of classification and localization makes object detection more complex than simple image classification, as the model needs to precisely identify multiple objects of potentially different classes and determine their positions within a single image.

A major difficulty in object detection lies in the complexity of real-world images, where multiple objects can appear at once, each differing in size, shape, angle, and possibly being partially hidden or surrounded by visual clutter. The model must correctly distinguish and locate each of these, often under less-than-ideal conditions.

To train such models effectively, the total loss used during learning typically blends two types of losses. Classification loss, which evaluates how accurately the model predicts the object classes (commonly using cross-entropy). Regression loss, which evaluates how precisely the model predicts bounding box coordinates, often using smooth L1 loss for better stability and handling of outliers. This combined loss encourages the model to both recognize what the object is and pinpoint where it is.

Image segmentation refers to the process of partitioning an image into distinct segments, each representing a specific object or region within the scene. This task is typically divided into two main categories: semantic segmentation and instance segmentation.

Semantic segmentation involves assigning a class label to every pixel in the image such that all pixels associated with a particular class are uniformly labeled. The primary objective of semantic segmentation is to achieve a dense, pixel-level classification across the entire image, thereby enabling detailed scene understanding.

Instance segmentation builds upon semantic segmentation by not only assigning a class label to each pixel but also distinguishing between individual instances of the same class. In other words, while semantic segmentation classifies all pixels belonging to a particular category collectively, instance segmentation ensures that each object instance within a class (e.g., two distinct vehicles) is uniquely identified at the pixel level. This approach is particularly beneficial in complex visual scenes involving multiple, closely positioned objects of the same category, enhancing the granularity and precision of object recognition

2.2.2 Traditional Feature Extraction

In computer vision, traditional feature extraction refers to the manual design of algorithms that identify and extract specific visual characteristics from images such as edges, textures, shapes, or other semantic attributes. These features, crafted based on domain expertise and image processing knowledge, serve as input to traditional machine learning models for tasks such as classification or object detection. For instance, hand-crafted features might emphasize object boundaries or surface patterns and are engineered in a way that facilitates their interpretation by learning algorithms.

A notable advantage of feature extraction is its effectiveness in scenarios with limited data. Because the features are predefined and focused on key visual cues, models can achieve acceptable performance without requiring extensive training datasets. This makes the method particularly suitable in domains where labeled data is scarce or costly to obtain.

Nevertheless, this approach encounters significant limitations when applied to large scale or highly variable datasets. Manually designed features often struggle to encapsulate the rich diversity and subtle nuances present in real-world imagery. As visual complexity increases, such features may fail to generalize to new data, particularly when confronted with variations in object scale, orientation, or illumination.

Moreover, the process of designing and optimizing features becomes increasingly inefficient as the complexity of the task escalates. Consequently, more advanced methodologies become necessary for achieving robust performance.

2.2.3 Convolutional Neural Networks

In the field of deep learning, CNNs are among the most widely used and influential models. Their primary advantage over earlier models is their ability to automatically identify relevant features without requiring human supervision. CNNs have found extensive applications in domains such as computer vision, speech processing, and facial recognition.

The structure of CNNs is inspired by the organization of neurons in the human and animal visual cortex. Just as the brain processes visual information through a hierarchical sequence of cells, CNNs simulate this process through layers of convolutional filters. Three key advantages distinguish CNNs from traditional fully connected networks are sparse interactions, parameter sharing and equivariant representations.

Unlike conventional neural networks, which require a fully connected structure, CNNs utilize shared weights and local connections, particularly suited for processing two-dimensional input data such as images. This architecture significantly reduces the number of parameters, simplifying the training process and improving computational efficiency. The local receptive fields in CNNs mimic the behavior of neurons in the visual cortex, focusing on small regions of an image at a time rather than processing the entire scene simultaneously. This localized feature extraction enables CNNs to capture spatial patterns more effectively, making them highly efficient for tasks involving image and signal analysis.

A widely used variant of CNNs, structurally similar to the multi-layer perceptron, consists of multiple convolutional layers followed by subsampling (pooling) layers, with fully connected layers at the final stages.

In a CNN, the input for each layer is structured in three dimensions: height, width, and depth, where the depth corresponds to the number of channels. For instance, in an RGB image, the depth is three, representing the red, green, and blue channels. Each convolutional layer contains multiple filters (kernels), which also have three dimensions, with width and height smaller than the input and depth equal to that of the input. These filters, which also have trainable parameters of weights and biases, generate feature maps by performing local convolutions with the input. The convolution operation involves computing a dot product between the input and the filter weights over small regions of the image.

Mathematically the operation in a convolutional layer is expressed as:

$$y(i,j) = \sum_{c=0}^{C-1} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x_c(i+m, j+n) \cdot w_c(m,n) + b$$

where:

- $y(i,j)$ is the output at spatial location (i,j),
- $x_c(i+m, j+n)$ is the input value at channel c and location $(i+m, j+n)$,
- $w_c(m,n)$ represents the filter weights for channel c,
- b is the bias term added after convolution,
- C is the number of input channels,
- M and N are the height and width of the filter.

The result is then passed through a non-linear activation function, such as the Rectified Linear Unit (ReLU) or the Sigmoid Linear Unit (SiLU). Activation functions are essential components in neural networks, allowing them to learn and represent complex patterns in the input data. By enabling the network to capture intricate relationships beyond linear associations, activation functions significantly enhance the model's expressiveness and its ability to generalize effectively to previously unseen data. These activation functions also accelerate learning, particularly in deep architectures, by mitigating vanishing gradient issues.

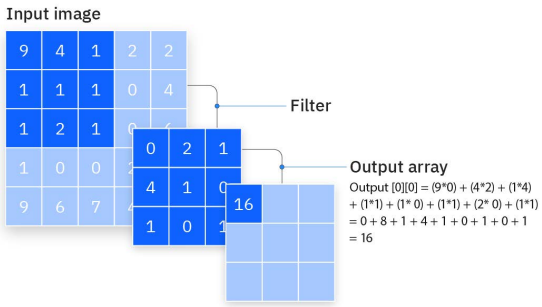


Figure 1: Example of the convolution operation.

The ReLU activation function is defined as:

$$f(x) = \max(0, x)$$

The SiLU, also known as the Swish activation function, is defined as:

$$f(x) = x \cdot \sigma(x)$$

where $\sigma(x)$ is the sigmoid function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Following the convolutional layers, down-sampling (pooling) layers reduce the spatial dimensions of the feature maps. This step decreases the number of parameters, accelerates training, and mitigates the risk of overfitting. The pooling function—commonly max pooling or average pooling—is applied to small regions of size $p \times p$ where p is the pooling kernel size. Max pooling selects the maximum value from the region while average pooling computes the mean value.

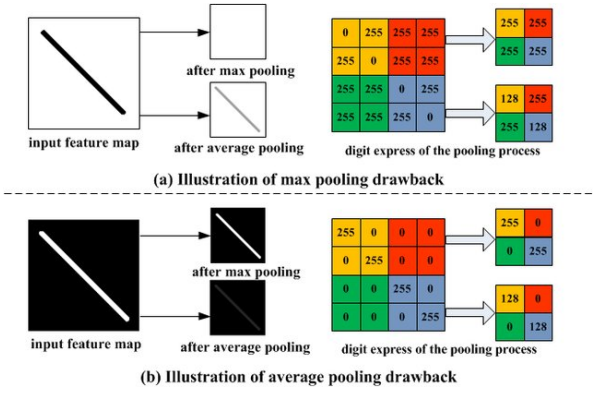


Figure 2: Examples of max pooling and average pooling and their drawbacks [41].

Finally, fully connected layers process the extracted features, transforming mid- and low-level representations into high-level abstractions. The final classification layer, often using softmax or support

vector machines, assigns a probability score to each class, indicating the likelihood that the given input belongs to a specific category.

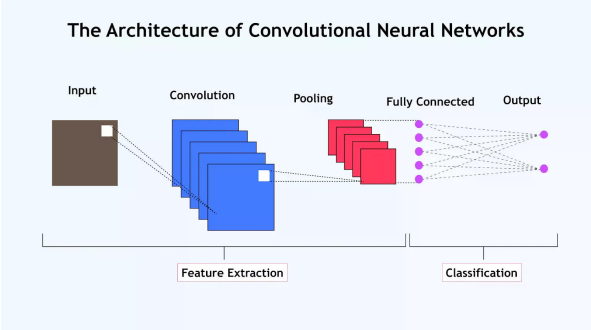


Figure 3: Example of a simplified CNN architecture.

The learning process in CNNs entails optimizing parameters—namely, weights w and biases b through backpropagation combined with gradient-based optimization algorithms. A loss function, typically cross-entropy for classification tasks, is used to quantify the model's prediction error. The gradients of this loss function with respect to each parameter ($\frac{\partial \mathcal{L}}{\partial w}$ and $\frac{\partial \mathcal{L}}{\partial b}$) are calculated and used to adjust the parameters in the direction that minimizes the loss.

Popular optimization algorithms include Stochastic Gradient Descent, RMSprop, and Adam. For instance, the Adam optimizer updates weights and biases as follows:

$$w \leftarrow w - \eta \cdot \frac{m_t}{\sqrt{v_t} + \epsilon}, \quad b \leftarrow b - \eta \cdot \frac{m_t}{\sqrt{v_t} + \epsilon}$$

Here, m_t is the exponential moving average of the gradient (first moment), v_t is the exponential moving average of the squared gradient (second moment), η is the learning rate and ϵ is a small constant to prevent division by zero. This approach helps stabilize training and accelerates convergence, especially in high-dimensional parameter spaces.

One outstanding characteristic of CNNs lies in their parameter sharing mechanism. Unlike fully connected networks where each neuron has a unique set of weights, CNNs apply the same filter (or kernel) across multiple spatial locations in the input image. This reuse of parameters significantly reduces the number of trainable weights, leading to greater computational efficiency and lower risk of overfitting, especially when training data is limited.

Additionally, CNNs exhibit spatial invariance, meaning they can recognize features or patterns regardless of their position within the image. This is achieved through convolution and pooling operations, allowing CNNs to effectively handle variations in object location and scale—an essential property for tasks such as object detection, image classification, and localization.

Modern CNN architectures integrate several advanced techniques to enhance learning efficiency, generalization capability, and training stability.

Batch normalization Batch normalization is widely employed to normalize the input distributions of each layer, thereby reducing internal covariate shift and facilitating faster convergence during training.

Dropout Dropout, a regularization method, randomly deactivates a subset of neurons during each training iteration, which helps to prevent overfitting and improves model robustness.

Residual connections Residual connections, prominently featured in architectures such as ResNet, enable the training of very deep networks by allowing gradients to bypass certain layers via identity mappings, thus mitigating the vanishing gradient problem.

Padding Padding, typically involving the addition of zero values around the image borders, is used to preserve spatial dimensions after convolution operations, ensuring that feature maps maintain a consistent size.

Strided convolutions Strided convolutions serve as an alternative to pooling for downsampling feature maps by applying convolutional filters with a stride greater than one, thereby reducing spatial dimensions while maintaining learnable parameters.

2.3 YOLO

YOLO is a prominent real-time object detection framework that performs both object detection and classification in a single pass through a neural network. Unlike traditional region-based detection methods that rely on separate stages for proposal generation and classification, YOLO treats object detection as a regression problem, directly predicting bounding boxes and associated class probabilities from the entire image. The architecture is typically composed of three core components: the backbone, which is responsible for extracting rich and high-level visual features from the input image; the neck, which enhances and fuses multi-scale feature representations to improve detection of objects at different sizes; and the head, which outputs the final predictions, including object class labels and bounding box coordinates. YOLO's unified design enables high inference speed and has made it highly suitable for applications requiring real-time detection, although earlier versions initially sacrificed some accuracy in favor of speed.

YOLOv11 is specifically designed to address the challenges associated with small object detection, offering enhanced accuracy without compromising the real-time inference speed that characterizes the YOLO family of models. Through architectural refinements and improved feature representation, YOLOv11 demonstrates superior performance in identifying small and densely packed objects, making it highly effective for applications requiring precise detection under constrained latency conditions. This iteration introduces enhancements to both the backbone and neck components of the architecture, significantly improving its feature extraction capabilities and enabling more precise object detection. In addition, the model incorporates optimized training pipelines that contribute to faster convergence

and improved processing speeds. These improvements make it well-suited for a broad range of deployment environments, including resource-constrained edge devices and scalable cloud platforms, thereby broadening its applicability in real-world, latency-sensitive applications.

2.3.1 YOLOv11 Architecture

Backbone The backbone constitutes a fundamental component of the YOLO architecture, tasked with extracting hierarchical features from the input image across multiple scales. This is achieved through a series of stacked convolutional layers and specialized computational blocks, which progressively capture both low-level and high-level representations of visual information. By generating feature maps at varying resolutions, the backbone enables the detection system to identify objects of different sizes and complexities, forming the foundational basis for accurate object localization and classification in subsequent stages of the network.

Neck The neck component of the YOLOv11 architecture plays a pivotal role in aggregating and refining feature representations across multiple scales before they are passed to the detection head. This module typically involves a combination of upsampling and concatenation operations that fuse feature maps from different layers of the backbone. By integrating low-level spatial details with high-level semantic information, the neck enables the model to effectively capture and utilize multi-scale contextual information. This is particularly beneficial for detecting objects of varying sizes and shapes within the same image, thereby enhancing the overall detection performance and robustness of the model.

Head The head of the YOLOv11 architecture is the final component in the detection pipeline, tasked with producing the ultimate predictions for object localization and classification. It takes the refined, multi-scale feature maps from the neck and processes them through convolutional layers specifically designed to predict bounding box coordinates, objectness scores, and class probabilities. The head operates across multiple spatial resolutions to enhance the model's ability to detect both large and small objects within the same image.

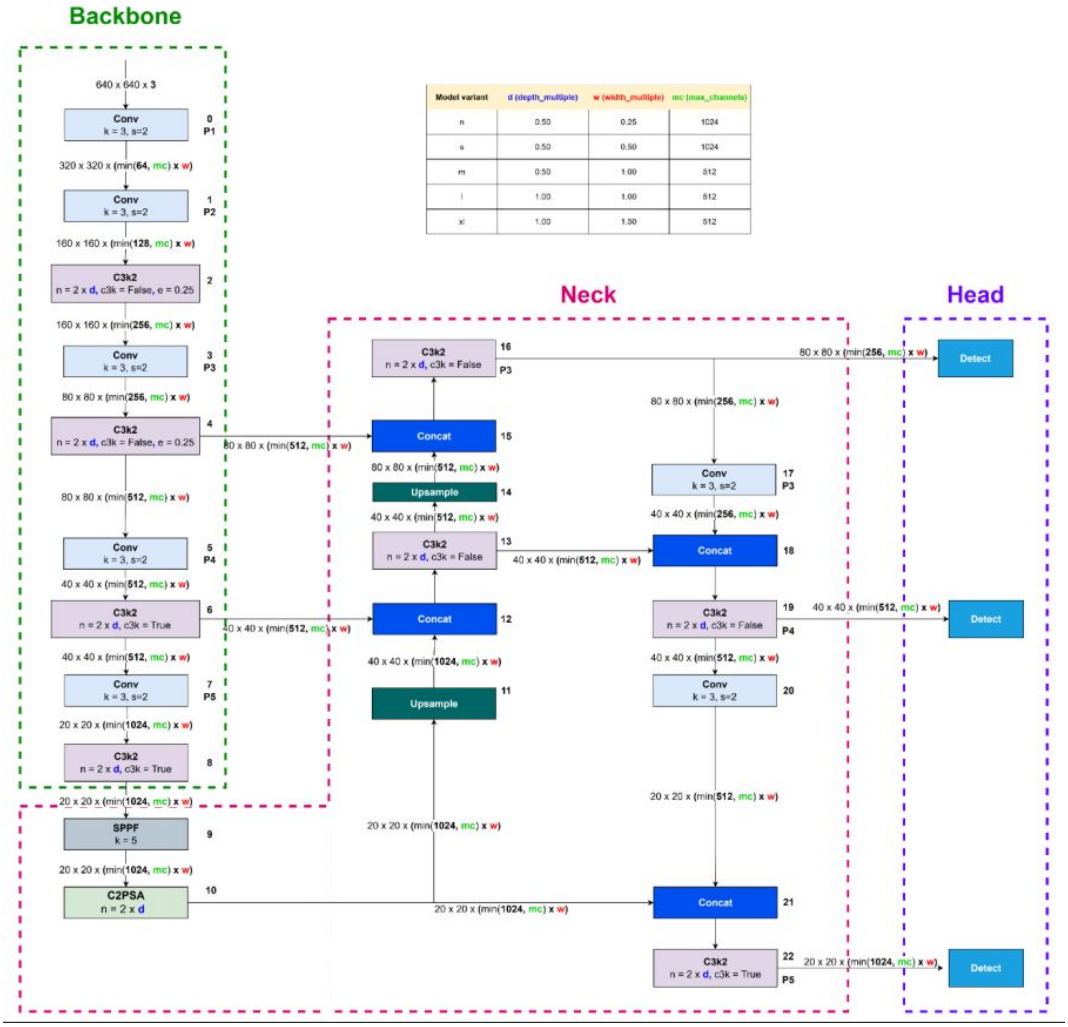


Figure 4: YOLOv11 Architecture [42].

Building Blocks of YOLOv11 YOLOv11 introduces several architectural innovations that significantly improve its ability to extract and process visual features, thereby advancing the overall performance of object detection tasks. Among the most notable enhancements are the inclusion of the C3k2 block, the Spatial Pyramid Pooling – Fast (SPPF) module, and the Cascaded Partial Self-Attention (C2PSA) attention mechanism. These components collectively optimize the model's computational efficiency and enhance its capacity to capture multiscale and complex features across varying object sizes and contexts.

The C2PSA module in particular, brings sophisticated spatial attention capabilities into the detection pipeline. By enabling the model to dynamically prioritize salient regions of an image, C2PSA enhances the detection of objects that are small, partially occluded, or embedded in cluttered backgrounds—scenarios that are typically challenging for conventional models. This targeted focus improves both localization and classification accuracy, making YOLOv11 particularly well-suited for demanding real-world applications such as medical imaging and autonomous systems.

1. Convolutional Block (Conv Block): The convolutional block is the most fundamental one in all YOLO models and consists of three layers: Conv2d, BatchNorm2d, and the SiLU activation function.

- **Conv2d Layer:** This layer performs convolution, sliding a kernel over the input data to produce a feature map. Its parameters include:
  - k: The number of filters applied during convolution, directly influencing the depth of the resulting output.
  - s: The stride, defines how far the kernel moves with each step, impacting the resolution of the output.
  - p: Padding refers to the addition of zero-valued borders around the input image to preserve its original spatial size after convolution.
  - c: denotes the number of input channels, which typically corresponds to color channels in an image (e.g., 3 for RGB images).
- **BatchNorm2d Layer:** This layer is used to normalize the activations from the preceding Conv2d layer. It standardizes the input features across the mini-batch by adjusting and scaling the activations, which helps in maintaining stable distributions of intermediate outputs throughout training. By doing so, BatchNorm2d mitigates internal covariate shift, accelerates convergence, and allows for higher learning rates. Moreover, it introduces a slight regularization effect, reducing the need for other regularization techniques such as dropout.
- **SiLU Activation Function**

YOLOv11 introduces two novel convolutional blocks in addition. The first is a convolutional block without an activation function, which is strategically placed to preserve linear transformations and facilitate gradient flow during backpropagation. The second is the DWConv, which significantly reduces computational complexity by applying a single convolutional kernel per input channel, rather than performing convolutions across all channels as in standard convolutions. This approach reduces the number of parameters and operations, making the network lighter and faster. Importantly, the kernel in each channel of the DWConv block has distinct matrix values, allowing the network to learn diverse spatial features from different input channels, thus maintaining representational richness while enhancing efficiency.

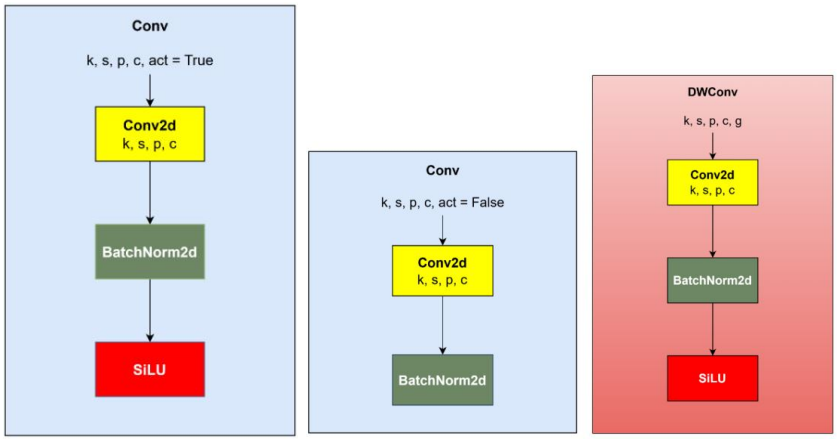


Figure 5: The standard YOLO Convolutional Block along with the two new YOLOv11 Blocks.

2. Bottleneck Block: The Bottleneck Block is an architectural component designed to improve computational efficiency and model depth in CNNs. It incorporates a shortcut (or residual) connection that allows the input to bypass the convolutional layers, directly connecting to the block's output. This mechanism facilitates better gradient flow during backpropagation, effectively addressing the vanishing gradient problem and enhancing the training of deep networks. When the shortcut is enabled, it introduces an identity mapping, allowing the network to preserve information and focus on learning residual functions. Conversely, if the shortcut is not used, the input undergoes transformation through two sequential Conv Blocks. The bottleneck design allows stacking of deeper layers while maintaining manageable computational costs, thereby enabling the network to learn more complex and expressive feature representations.

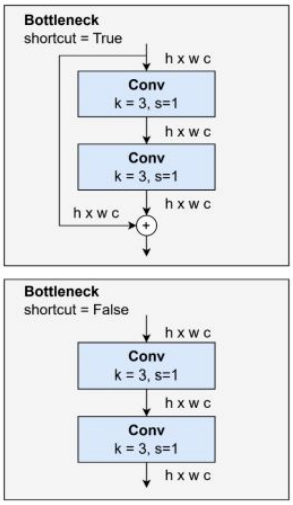


Figure 6: The bottleneck blocks.

3. C3k Block: The C3k block is a modular structure within the YOLOv11 architecture, composed of three convolutional blocks and multiple bottleneck units. These bottlenecks are configured to either

include a shortcut connection (shortcut=true) or not (shortcut=false).

4. C3k2 Block: The C3K2 block employed in YOLOv11 represents a refined evolution of the Cross Stage Partial (CSP) bottleneck architecture seen in earlier YOLO versions. This module is designed to optimize information flow and enhance computational efficiency by splitting the input feature map into separate parts and applying a sequence of smaller kernel convolutions. These smaller kernels are not only faster to compute but also reduce the number of learnable parameters, making the network more lightweight without compromising performance. By processing sub-feature maps independently and merging them post-convolution, the C3K2 block achieves richer feature representation and greater diversity in learned patterns. It is particularly adept at learning multiscale features, leveraging feature vector switching and multi-layer convolutions to expand the network's receptive field. This allows the architecture to better capture complex visual patterns and detect objects of varying sizes and scales, making C3K2 a critical component in YOLOv11's improved detection accuracy and efficiency.

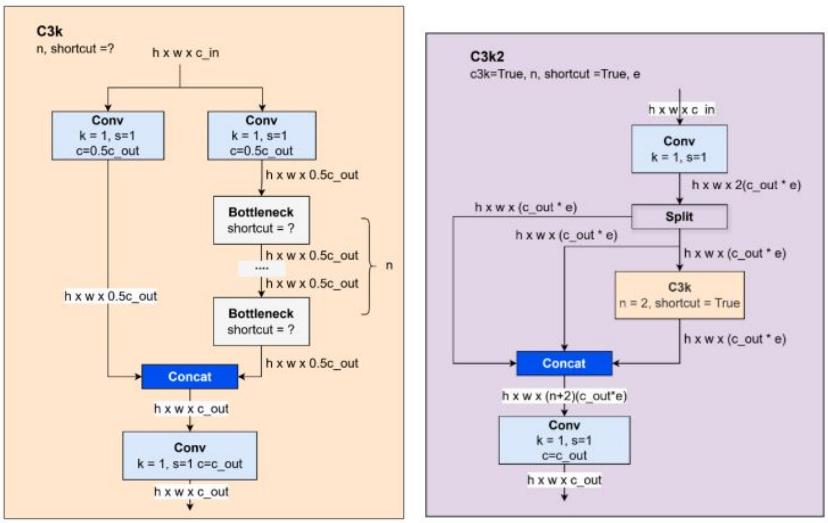


Figure 7: The C3k2 Block.

5. SPPF Block: SPPF is a module designed to enhance the representational power of CNNs by capturing multi-scale contextual information within feature maps. It achieves this by applying max pooling operations at different scales, which enables the network to aggregate features from varying receptive fields and thus learn both fine-grained and high-level abstractions. SPPF is an optimized version of the original Spatial Pyramid Pooling (SPP) used in earlier YOLO architectures. While traditional SPP utilizes multiple max pooling kernel sizes (e.g., 3, 5, and 9), SPPF simplifies this structure by employing a single kernel size of 5, applied sequentially rather than in parallel. This design choice significantly reduces computational complexity and the number of floating-point operations while maintaining robust multi-scale feature extraction capabilities. As a result, SPPF offers a more computationally efficient solution for capturing spatial hierarchies in the feature maps, significantly improving the network's ability to capture especially small objects, which has been a

challenge for earlier YOLO versions.

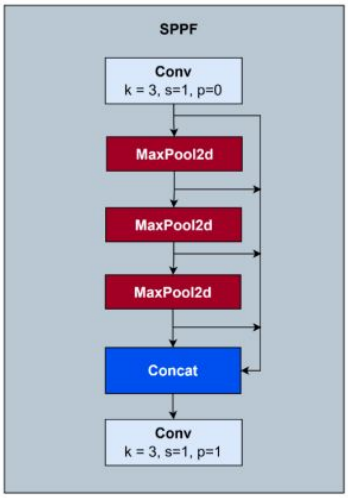


Figure 8: The SPPF Block.

6. PSA and C2PSA Blocks: The attention module in modern object detection architectures serves the purpose of modeling relationships between pixels within an image or video frame, allowing the network to focus on the most informative regions. The Partial Self-Attention (PSA) module is composed of a single attention mechanism followed by a feed-forward layer. The attention layer computes relationships between spatial locations, enabling the model to selectively emphasize informative regions. Following this, the feed-forward layer further refines the attended features, enhancing the network's capacity to learn meaningful patterns. Despite its simplicity, this streamlined structure allows the PSA to efficiently capture spatial relationships within the feature maps by focusing on a partial subset of the input, offering a good balance between computational efficiency and representational power.

The C2PSA module introduces a deeper and more refined structure by stacking multiple PSA operations. It begins with a convolutional block that divides the input into two branches: one branch is forwarded directly to a concatenation block, while the other undergoes further processing through an attention mechanism followed by another convolutional block. The outputs from both branches are then concatenated and passed through an additional convolutional layer to fuse the information. This design allows for the iterative refinement of attention, capturing more complex and nuanced interdependencies among features. Despite its increased complexity, C2PSA maintains computational efficiency, enabling it to learn richer representations without incurring significant additional computational cost. The integration of spatial attention mechanisms enhances the model's ability to focus on salient regions within the image, which is particularly beneficial for tasks involving small or occluded object detection. This makes C2PSA a powerful addition.

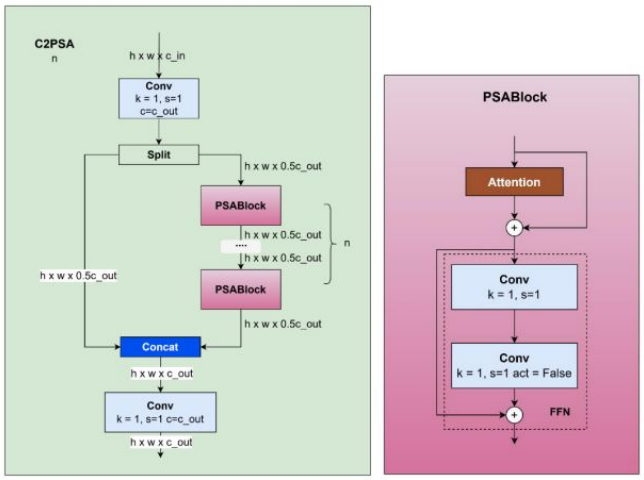


Figure 9: C2PSA and PSA blocks.

7. Detect Block: The detect block plays a crucial role in finalizing predictions and comprises two distinct components: the classification head and the regression head. The classification head is responsible for predicting the class probabilities of detected objects, while the regression head predicts the bounding box coordinates. These two heads are architecturally distinct due to differences in computational demands. Specifically, the classification head may be structurally larger, potentially twice the size of the regression head, to handle the more complex task of multi-class probability estimation. However, the regression head holds greater influence over the overall detection performance, as accurate localization is critical. To mitigate the increase in computational cost and model complexity introduced by the larger classification head, YOLOv11 integrates depthwise convolutional (DWConv) blocks. These blocks streamline the classification head by performing efficient channel-wise convolutions, substantially reducing the number of parameters and floating-point operations without compromising accuracy.

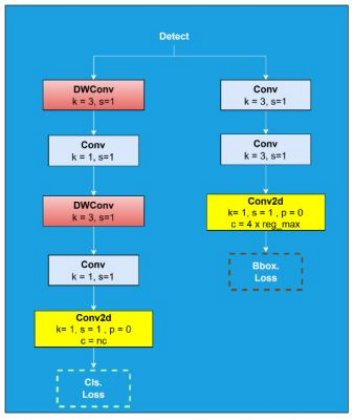


Figure 10: The Detect Block.

3 Related Work

Computer vision has transformed numerous domains by enabling automated systems to interpret and analyze visual data with high precision. Its applications span across fields such as industrial automation, environmental surveillance, and medical imaging. With ongoing advancements in image processing, object detection, and pattern recognition, computer vision technologies are now capable of addressing increasingly complex visual tasks with greater accuracy and efficiency.

In the context of veterinary parasitology, computer vision is emerging as a powerful tool for aiding microscopic diagnosis. One particularly promising application lies in the detection of intestinal parasites in animal fecal samples—a task traditionally dependent on time-intensive manual examination under a microscope. Accurate identification of parasites is crucial for diagnosing infections and guiding appropriate treatment strategies, especially in cases where morphological differences between species are subtle.

The integration of computer vision with digital microscopy enables automated identification and localization of parasitic organisms within complex image backgrounds. These systems offer the potential to significantly reduce diagnostic time, standardize results across operators, and enhance diagnostic accuracy.

This section reviews the current state of computer vision technology, highlighting its general applications and recent advancements. In particular, it focuses on the specific application of computer vision in microscopic parasite detection, discussing how these technologies are advancing the field of veterinary diagnostics and improving the efficiency and accuracy of parasite identification in clinical and laboratory settings.

3.1 State of the Art Computer Vision Applications

Object Detection The field of object detection has undergone substantial advancements over time. Traditional methods, such as the Viola-Jones detector [43] and Deformable Part-based Models (DPMs) [44], were grounded in sliding window strategies and the use of hand-crafted features, including Haar wavelets [45] and Histograms of Oriented Gradients (HOG) [46]. These techniques relied on scanning the image with fixed-size windows at various scales to localize objects. Although effective in constrained environments, these early approaches often exhibited limited generalization capabilities, particularly in complex scenes involving diverse object scales and aspect ratios [47].

The emergence of deep learning marked a pivotal shift in object detection methodologies. One of the foundational contributions in this new paradigm was the Region-based Convolutional Neural Network (R-CNN) [48], which utilized selective search to generate region proposals followed by CNN-based classification for each candidate region. Despite its accuracy, R-CNN suffered from inefficiency due to the need to process each region independently through a CNN.

To overcome this limitation, Fast R-CNN [49] introduced a more efficient framework that leveraged a shared convolutional feature map for both region proposals and classification, greatly accelerating

inference. Building on this, Faster R-CNN [50] integrated a Region Proposal Network (RPN) directly into the architecture, enabling real-time generation of object proposals and significantly enhancing both speed and detection accuracy.

Subsequently, Mask R-CNN [51] expanded the capabilities of Faster R-CNN by introducing a parallel branch for pixel-level instance segmentation, enabling simultaneous object detection and segmentation.

Advancements in Real-Time Object Detection: One-Stage Detectors

In response to the growing demand for real-time object detection, one-stage detection architectures were developed to optimize inference speed without relying on region proposal mechanisms. A seminal contribution in this area was the YOLO framework [39], which pioneered real-time detection by employing a single CNN to concurrently predict object bounding boxes and their associated class probabilities. While YOLO achieved significant gains in computational efficiency, this came at the expense of reduced accuracy when compared to two-stage detectors.

To address the trade-off between speed and accuracy, the Single Shot MultiBox Detector (SSD) [52] was proposed. SSD utilized feature maps at multiple scales to detect objects of varying sizes, thereby improving localization performance and achieving a more favorable balance between detection speed and accuracy.

Subsequently, RetinaNet [53] further advanced one-stage detectors by introducing Focal Loss, a novel loss function designed to address the severe foreground-background class imbalance inherent in dense object detection. This innovation led to substantial improvements in detection accuracy while preserving the high inference speed characteristic of one-stage architectures.

Transformer-Based Advances in Object Detection

The integration of transformers into object detection marked a significant shift from conventional convolutional paradigms. The Detection Transformer (DETR) [54] introduced a novel end-to-end architecture that eliminated the need for hand-crafted region proposals and complex post-processing by leveraging the global attention mechanism of transformers. Despite its conceptual elegance and strong detection capabilities, DETR suffered from slow convergence during training, which limited its practical applicability.

To overcome this, Deformable DETR [55] introduced deformable attention modules that focused computational resources on a sparse set of key spatial locations. This innovation led to substantial improvements in training efficiency and detection performance. Concurrently, the YOLO series continued to evolve, with YOLOv7 [56] refining the architecture to achieve superior accuracy while maintaining real-time inference capabilities. YOLOv8 [57], the most recent iteration, introduced enhanced model scaling strategies and improved feature representations, solidifying its position as a state-of-the-art solution for real-time object detection.

In parallel, DINO (DETR with Improved Training Optimization) [58] addressed key training in-

efficiencies in the original DETR, resulting in more robust performance across various vision tasks. Collectively, these advancements underscore the growing synergy between transformer-based architectures and real-time object detection.

Further advancements in object detection include YOLO-NAS [59], which leverages neural architecture search (NAS) to automatically design optimized YOLO-based architectures. This approach has achieved state-of-the-art results across multiple object detection benchmarks by balancing accuracy and computational efficiency. Additionally, the Segment Anything Model (SAM) [60], developed by Meta AI, though primarily designed for segmentation tasks, has exhibited remarkable zero-shot object detection capabilities. Its ability to generalize across a wide range of visual domains without task-specific training marks a significant step forward in model versatility and adaptability.

These developments highlight the ongoing evolution of object detection methodologies—moving from manually designed models toward automatically optimized and generalizable architectures. This progression is pushing the limits of what is possible in terms of detection accuracy, efficiency, and scalability across various real-world applications.

Image classification Image classification, as one of the foundational pillars of computer vision, has witnessed a dramatic evolution over the past decade, transitioning from relatively shallow CNNs to highly efficient and transformer-based architectures. The breakthrough moment came with the introduction of AlexNet [61] in 2012, which demonstrated the potential of deep learning in large-scale image classification tasks by securing a landmark victory in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The architecture utilized deep convolutional layers, ReLU activations, and dropout regularization, thereby setting the precedent for future models. However, AlexNet also revealed the high computational demands required for training deep networks, prompting a subsequent wave of architectural innovations.

Following AlexNet, VGGNet [62] introduced a more uniform and deeper architecture, relying on small convolutional kernels throughout. Although it offered improved accuracy by increasing the depth of the network, VGGNet also came with the trade-off of significantly higher parameter counts and memory usage, making it less suitable for deployment in resource-constrained environments. To address this, GoogleNet (Inception) [63] introduced the concept of parallel convolutional layers of multiple receptive fields within the same block, creating what became known as the Inception module. This architecture improved computational efficiency without compromising model performance by adaptively capturing multi-scale spatial information.

Residual Networks A pivotal moment came with the introduction of ResNet [64], which tackled the growing challenge of training very deep networks by introducing residual connections—shortcuts that allow gradients to bypass certain layers, effectively mitigating the vanishing gradient problem. ResNet enabled the training of networks with depths exceeding 100 layers, dramatically improving performance on image classification benchmarks and establishing a new standard for deep learning

models. The concept of residual learning has since become a core architectural principle across many modern computer vision models.

The increasing demand for lightweight models capable of running on edge devices led to the development of architectures like MobileNetV2 [65], which introduced inverted residuals and linear bottlenecks to reduce computational complexity while preserving accuracy. This was particularly transformative for mobile applications, as it enabled real-time image recognition on smartphones and embedded systems. Following this trend of balancing efficiency and performance, EfficientNet [66] proposed a principled compound scaling method that simultaneously scales a model's width, depth, and input resolution. This approach achieved state-of-the-art results across multiple benchmarks using fewer parameters and FLOPs compared to its predecessors.

EfficientNetV2 [67] further refined this scaling strategy by incorporating improved training dynamics, faster convergence, and better utilization of training resources. These innovations paved the way for more sustainable and scalable deployment of deep learning models, particularly in scenarios where both accuracy and resource-efficiency are crucial.

Transformer-Based Classification More recently, the introduction of transformer-based architectures marked a paradigm shift in image classification. The Vision Transformer (ViT) [68] was the first model to directly apply the self-attention mechanism—originally developed for natural language processing—to sequences of image patches. By treating image patches as tokens and processing them with transformer encoders, ViT captured long-range dependencies that were traditionally difficult to model using convolutional operations. Despite its remarkable performance on large-scale datasets, ViT lacked some of the inductive biases inherent to CNNs, such as locality and translation equivariance, and thus required extensive data and computational power to train effectively.

In response to these limitations, DeiT (Data-efficient Image Transformer) [69] introduced optimization strategies such as knowledge distillation to train transformers with less data, making them more accessible for real-world applications. Swin Transformer [70] advanced the transformer architecture further by implementing a hierarchical structure with shifted windows, allowing the model to capture both local and global contexts while maintaining linear computational complexity with respect to image size. This made Swin Transformer particularly effective in a variety of vision tasks, including object detection and semantic segmentation, in addition to classification.

Hybrid Classification Models At the intersection of CNNs and transformers, ConvNeXt [71] emerged as a modern reinterpretation of convolutional models, incorporating design elements inspired by transformers—such as large kernel sizes, layer normalization, and GELU activations—while retaining the computational efficiency and intuitiveness of convolutional operations. ConvNeXt demonstrated that with proper architectural tuning, CNNs could still compete with or outperform transformer-based models on standard benchmarks.

Additionally, hybrid architectures such as convolution-and-attention-based models (CoAtNet) [72]

combined the best of both worlds by integrating convolutions and self-attention mechanisms in a unified framework, achieving superior generalization and robustness. Lightweight models like MobileViT [73] pushed this even further by delivering transformer-level performance on mobile hardware, reinforcing the trend toward democratizing AI technologies.

Self-supervised learning approaches also gained prominence, particularly with the introduction of Masked Autoencoders (MAE) [74], which extended the masked language modeling paradigm from NLP to vision. By randomly masking image patches and training the model to reconstruct them, MAEs enabled the learning of rich visual representations with minimal labeled data. Finally, models such as EVA [75] showcased the scalability of vision transformers, achieving unprecedented results across a wide array of vision benchmarks, from classification to dense prediction tasks.

Collectively, these advancements have not only pushed the boundaries of image classification but also laid the groundwork for modern computer vision applications in domains as diverse as healthcare, autonomous driving, and digital pathology. They represent a progressive convergence of accuracy, efficiency, and flexibility, continually shaping the way visual data is processed and understood by machines.

3.2 Intestinal Parasite Detection

There are many articles that describe the analysis of microscopy images of different parasites. The majority of them are related to human parasites, their detection, and identification. Several of these studies on human parasites, especially more recent ones, make use of large, high quality and more importantly open datasets. Computer vision applications for the detection of canine and feline intestinal parasites specifically, are scarce in comparison. Datasets on this subject being mostly private is a critical factor which makes the comparison of existing study results challenging. For these reasons, there is value in covering previous works regarding human parasites, as we can get a better grasp of what models and techniques have been experimented with and how they compare with each other. The same principals should apply to experiments with data from dogs and cats, in general.

3.2.1 Studies on Human or Other Parasites

One of the first studies utilizing deep learning, proposed a method to automate routine fecal examinations for parasitic diseases using digital image processing techniques coupled with artificial neural networks (ANN). A dataset of 82 microscopic images featuring seven common human helminth eggs was utilized. Species identification was conducted using an ANN classification system composed of two multilayer perceptron (MLP) neural networks. The first network (ANN-1) was responsible for distinguishing parasite eggs from artifacts, ensuring that only parasitic organisms are isolated from other similarly sized elements present in fecal samples. The second network (ANN-2) then classified the species of the detected eggs identified by the ANN-1, utilizing the same numerical features for classification. The performance of the ANNs was assessed using ten fold cross-validation to minimize

bias from training sample selection. Results demonstrated an average detection accuracy of 84% for the ANN-1 and an average classification accuracy of 83% for the ANN-2. [76].

A similar early work also made use of a MLP-based system for detecting *Cryptosporidium parvum* oocysts, aiming to streamline analysis time and improve diagnostic accuracy. The system was trained on a dataset comprising 525 images, including labeled oocyst images and non-oocyst images. The system correctly identified authentic oocyst images with an accuracy ranging from 80% to 97%. For non-oocyst images, the accuracy ranged from 77% to 82% on the test dataset [77].

Finally on the experiments with MLPs, another study focused on analyzing images of the protozoa *Cryptosporidium spp.* and *Giardia spp.*. One MLP was developed for detecting *Cryptosporidium* oocysts was trained on a dataset of 1,586 images, while a different MLP for *Giardia* cysts was trained with 2,431 images. Both networks were validated using previously unseen datasets: 500 images for *Cryptosporidium* (comprising 250 positive and 250 negative samples) and 282 images for *Giardia* (including 232 positive and 50 negative samples). The system achieved an identification accuracy of 91.8% for *Cryptosporidium* oocysts and 99.6% for *Giardia* cysts [78].

Later on, a group of researchers explored the application of an Adaptive Neuro-Fuzzy Inference System (ANFIS) for classification tasks, integrating a moment-invariant-based feature extraction mechanism to enhance pattern recognition. Their study focused on developing a robust computational model that combines fuzzy logic and ANNs to handle nonlinear and uncertain data effectively. The proposed system followed a two-stage approach: first, feature extraction using moment invariants, and second, classification using ANFIS. In the first stage, the system applies moment invariants to extract shape-based features from input data. Moment invariants, such as Hu's moments, ensure that extracted features remain unchanged under transformations like rotation, scaling, and translation, making them highly suitable for pattern recognition. This preprocessing step reduces data dimensionality and enhances classification robustness by providing transformation-independent feature vectors. The second stage involves training and utilizing the ANFIS model for classification. ANFIS is structured as a five-layer hybrid system incorporating both fuzzy if-then rules and ANN learning to map input features to classification outputs. Their technique was applied to recognize 16 different human parasite eggs from their microscopic images. The ANFIS system achieved a correctness rate of 93.49%. However, the primary limitation was the misclassification of eggs with similar shapes, highlighting the need for further refinements to the technique [79].

Another advance was made with the Multi-Class Support Vector Machine (MCSVM) classifier, which was evaluated in a study involving 16 human parasites and achieved a success rate of 97.70% [80]. This method, like others, followed four stages: preprocessing, feature extraction, classification, and testing. In the preprocessing stage, digital image processing techniques such as noise reduction, contrast enhancement, thresholding, and morphological and logical operations were applied. During feature extraction, the invariant moments of the preprocessed parasite images were computed. The MCSVM classifier was then employed to classify the features. However, the system faced a recurring challenge—misclassification of eggs with similar shapes. Among these stages, preprocessing was

identified as the most critical for the method's success.

With the previous works in mind, one particular research group's efforts concentrated on the automatic segmentation and classification of microscopy images containing fecal impurities. Their initial work identified the 15 most prevalent species of protozoa, eggs, and helminth larvae found in Brazil. Comparative evaluations were conducted to assess the performance of optimum-path forest (OPF), ANN-MLP, and SVM classifiers, both with and without Bagging and AdaBoost, techniques used to construct classifier ensembles. Their dataset consisted of 1791 images containing intestinal parasites. Their analysis revealed that the OPF classifier was the most effective for the species studied, achieving 93.00% sensitivity, 99.17% specificity, and an average Kappa coefficient of 0.84 [81]. In a subsequent study of theirs, the researchers sought to answer the question of whether their previous system can benefit from the higher effectiveness of deep neural networks without compromising its efficiency. To this end, they developed a CNN with random kernels and utilized it for the extraction of feature vectors, which would then be fed into a linear SVM classifier. The authors reported significant gains in classification accuracy over their previous work in most of their tests, with an average Kappa coefficient of around 0.96 [82].

This next work is one of the first to propose a framework leveraging CNNs to evaluate performance across three distinct microscopy tasks: identifying intestinal parasite eggs in stool samples, detecting tuberculosis in sputum samples, and diagnosing malaria in thick blood smears. Experts annotated the images by marking bounding boxes around relevant entities. Approximately 1,000 images per task were used to train the model, which was subsequently tested on separate datasets. The results demonstrated that the CNN significantly outperformed traditional medical imaging techniques in all evaluated cases, marking a substantial improvement in diagnostic accuracy [32].

Experimenting further, researchers worked with microscopic images of stool samples which contained multiple parasites within a single image as well as numerous background elements. To address this problem, each parasite was individually isolated prior to recognition. This isolation was achieved through the process of image segmentation, specifically edge detection, which separates individual parasites from the background and other objects in the image. For the recognition of the parasites, principal component analysis (PCA) was employed for feature extraction and dimensionality reduction from the pixel data of the extracted parasite images. Classification was performed using a probabilistic neural network (PNN), a specialized form of radial basis neural networks, designed primarily for classification tasks. They serve as an alternative to traditional radial basis networks, offering distinct advantages such as ease of learning and instantaneous processing. These features make PNNs particularly effective for applications requiring rapid and straightforward model training. The developed algorithms were tested on a dataset of 900 microscopic images representing 15 different species of human intestinal parasites. The results demonstrated a 100% recognition success rate [83].

One particular study regarding intestinal parasites of various animal species had the goal of developing a cost-effective, automated parasite diagnostic system that eliminates the need for special sample preparation or trained personnel. The system utilized a trained CNN to automatically seg-

ment and analyze images, effectively distinguishing parasite eggs from background debris. The CNN followed a U-Net [84] architecture; however, unlike the original U-Net, the first max-pooling layer was repositioned to a deeper layer within the network. This adjustment enhanced the learned features' invariance to minor translations in the input, thereby improving the network's robustness and generalization capability. For the training set, a total of 872 images were utilized, comprising 564 images containing eggs and 308 without eggs. Additionally, 79 images with eggs were set aside as a test set to evaluate the model's performance. The labeled images correspond to species including monkeys, sheep, dogs, and cows. Postprocessing of the CNN output provides both species identification and egg counts. Validation against manual counts conducted by a trained operator demonstrated excellent performance, with overall accuracy rates of 92% to 96% for Eimeria species and 100% for nematodes. Notably, debris did not pose a challenge in this study, as the software was specifically trained to recognize and exclude it [85].

A following notable study focused on developing an automated system for detecting and classifying parasitic worm eggs in human stool microscopy images using the deep learning model Faster R-CNN [86]. This approach generates region proposals while simultaneously predicting object boundaries and classification scores for each identified region. The dataset was comprised of 246 images of 8 different parasite species. The model demonstrated high performance, achieving an mAP value of 97.67% [87].

Researchers acknowledged that staining plays a crucial role in improving the accuracy of automated parasite identification. To this end, a CNN model was trained to detect intestinal protozoa in human fecal samples stained with trichrome. The model was designed to be capable of screening out negative trichrome slides while flagging potential parasites for manual confirmation. The object detection model architecture was a three color channel CNN based upon SSD Inception v2 [88]. It was trained using conventional protozoa as distinct classes, with 1,394 to 23,566 exemplars per class. Despite limitations, such as the low prevalence of certain species, the study demonstrated that combining slide scanners with artificial intelligence software achieved a 98.88% positive agreement and a 98.11% negative agreement compared to manual microscopy [89].

One of the first papers to introduce YOLO-based solutions, investigated the application of YOLOv3 [90] for detecting helminth eggs in stool samples. The study utilized a dataset of 126 images, containing *Schistoma*, *Ascaris* and *Trichuris* eggs, captured using Android smartphones with varying resolutions and camera settings. Given the limited size of the dataset, data augmentation techniques such as rotations, flips, color adjustments, and shears were applied to enhance variability and improve model robustness. In terms of detecting helminth eggs, after training the tiny YOLO v3 model with 2-fold validation, the system achieved an average Intersection over Union (IoU) score of 0.8683 and an average sensitivity of 0.9105. Finally, the system achieved an average sensitivity of 0.8722 in terms of identifying the helminth eggs. The developed model functions in real-time, enabling automated detection and instant annotation of potential parasites, making it highly suitable for diagnostic applications [91].

The following study investigated the use of deep learning for recognizing *Ascaris lumbricoides* eggs, aiming to develop a prototype tool for parasite egg detection in medical diagnostics. The primary

objective was to accurately identify three distinct types of *A. lumbricoides* eggs: infertile eggs, fertile eggs, and decorticate eggs. The dataset for training and testing comprised images of *A. lumbricoides* eggs with a total of 600 images (200 per egg type). The experimental process consisted of two key phases. In the first phase, the researchers tested CNN models with one to ten convolutional layers, noting that increasing the layers beyond three reduces classification accuracy due to the loss of critical image features. In the second phase, they optimized the best-performing models by incorporating ReLU activation, max-pooling for feature reduction, and dropout to mitigate overfitting. The final CNN architecture employed a fully connected layer for classification, using precision, recall, and accuracy metrics to evaluate performance. Results indicated that the CNN model performs significantly better than traditional classification approaches. The confusion matrix showed that infertile eggs achieved 100% precision, fertile eggs 93.3%, and decorticate eggs 86.6%. The overall classification accuracy was 93.33%, demonstrating the robustness of the deep learning model in recognizing parasite eggs. However, the study acknowledged limitations such as manual parameter tuning, which could be improved through automated hyperparameter optimization in future research. [92].

An extensive work [93] evaluated the performance of two out-of-the-box fine-tuned models being AlexNet [61] and GoogleNet [63] in diagnosing intestinal parasite eggs in human stool samples, comparing their results to a custom-trained CNN [32] designed for the same task. The dataset included 6,500 image patches for AlexNet (10.9% positive), 6,461 for GoogleNet (11% positive), and 2,071 for the custom CNN (30.5% positive). Among the tested models, AlexNet performed the best, achieving a perfect ROC AUC score of 1.00 and an Average Precision of 0.93 on the test images. GoogleNet achieved an AUC of 0.99 and an average precision of 0.83. Finally, the custom-trained CNN achieved an AUC score of 1.00 and an average precision of 0.90. In all cases, the models required very few computing resources to run, indicating that out-of-the-box models have the potential to be used in real-world medical diagnostics.

Another study which evaluated and compared, at the time, state of the art pre-trained models, took place where the authors had to face the challenge of parasite detection, having to work exclusively with images captured using a low-cost USB microscope. Their dataset consisted of 162 total images, each containing 1-3 eggs of four genera of parasites. The proposed CNN models demonstrated robust capability in identifying key features of various parasitic eggs, even in low-quality images. Among the tested frameworks, ResNet50 [64] and AlexNet exhibited superior accuracy, outperforming SSD, and Faster R-CNN. These two models achieved testing accuracies of up to 96.93% and 98.25% for AlexNet and ResNet50, respectively. These promising results suggest that this method could be effectively applied in real-world fecal examinations using USB microscopes [94].

With the growing popularity of YOLO models, this next research proposed a detection system, based on YOLOv5, for identifying eight kinds of human parasite eggs in fecal microscopic medical images. A total of 281 sample images were used in this work, of which 236 were used for training and 45 for testing. Experimental evaluation revealed an mAP of 0.994 on the test dataset. Moreover, the system demonstrated great efficiency, processing each fecal microscopic image in under 25 milliseconds

using a GPU. In order to verify the effectiveness of this method, it was compared with other common object detection algorithms such as Faster-RCNN and RetinaNet [95]. The YOLOv5 based system proved superior in both accuracy and speed [96].

Continuing the trend, an extensive study was conducted, which utilized three YOLO-based approaches, YOLOv4-Tiny [97], YOLOv3 [90], and YOLOv3-Tiny, to identify protozoan cysts and helminthic eggs in human fecal samples. A dataset of 1,597 images prior to data augmentation, expanded to 182,058 images post-augmentation, was used to train the models for recognizing 34 classes of intestinal parasites. For evaluation, the models were tested using 176 images before augmentation and 4,752 images after augmentation, containing various intestinal parasitic objects. YOLOv4 employed the CSPDarknet53 [98] backbone as its feature extractor, replacing the Res block module with a combination of convolutional and CSP block modules. This design contributed to improved model efficiency and accuracy. The YOLOv4-Tiny model outperformed the other approaches, achieving the metrics of 96.25% precision, 95.08% sensitivity, 99.75% accuracy, 95.66% f1-score and an mAP of 0.963. These results highlight YOLOv4-Tiny's superior performance in detecting protozoan cysts and helminthic eggs, making it a promising tool for automated parasitological analysis [99].

Inspired by advancements in deep learning, another study adapted proven architectures to address the detection of human parasitic eggs in microscopic images. The proposed framework employed a two-step process. The first step used a Generative Adversarial Network (GAN) to enhance image quality, transforming low-resolution images into high-resolution ones. These pre-processed images were then passed to a Faster R-CNN model with a ResNet50 [64] backbone for object detection. The dataset included 2,907 images across five classes: Ascaris lumbricoides (558 images), Hookworm (550 images), Opisthorchis viverrine (549 images), Taenia spp. (551 images), and Trichuris trichiura (699 images). While the techniques yielded promising results, with very high precision and recal values, further refinements were required to improve detection for certain challenging egg types [100].

The following paper introduced a multi-modal learning detector designed to localize and classify human parasitic eggs into 11 distinct categories. The proposed framework leveraged EfficientNet [66], a family of models optimized for efficient classification, as the baseline architecture. EfficientNet serves as the backbone for the EfficientDet [101] framework, which extends its capabilities to detection and segmentation tasks. Experiments were conducted using the Chula-ParasiteEgg-11 dataset [102], comprising 11,000 microscopic images across 11 categories. The dataset was used to train both an EfficientDet model with an EfficientNet-v2 backbone [67] and to fine-tune a pre-trained EfficientNet-B7 model integrated with an SVM classifier. The multi-modal approach demonstrated robust performance, achieving an accuracy of 92% and an F1 score of 93% [103].

A similar study once again tackled the challenge of the Chula-ParasiteEgg-11 dataset, detecting human parasitic eggs. Several advanced CNN architectures were employed, such as High-Resolution Network (HRNet) [104], ResNet-101 [64], and ResNeXt-101 [105] as backbone models. The approach incorporated multi-task learning, where instance segmentation models were trained using pseudo-ground truth masks. These masks were initially produced by a class-agnostic segmentation model and

served as supplementary supervisory signals to enhance overall performance. To improve detection accuracy, outputs from both single-task and multi-task architectures were ensembled to identify parasitic egg cells and create pseudo-labels for the test set. Bounding box predictions from various models were merged using a weighted box fusion algorithm [106]. These pseudo-labels were then utilized in successive training rounds to refine the models. The results demonstrated that integrating these strategies yielded superior performance compared to relying on individual model predictions. The model's performance was assessed through metrics such as mAP score, mean IoU, and mean F1 score, achieving values of 0.956, 0.934, and 0.988, respectively [107].

In the following work, YOLOv5 was implemented in combination with variant cascadeRCNNs to detect human parasitic eggs in microscopic images. The proposed model was trained on the Chula-ParasiteEgg-11 dataset containing 11,000 images across 11 classes. Once all models had been tested, Weighted Boxes Fusion was utilized as an ensemble methods. This strategy allows the model to generate more precise object anchors with increased confidence. Additionally, it was observed that biologically inspired augmentation techniques significantly enhanced performance, highlighting the importance of task-specific data augmentation in biological applications. The model demonstrated strong performance, achieving a mAP of 0.952 and a mean IoU of 0.966 [108].

An important study compared the performance of various deep learning models, such as YOLOv8 [109], Detectron2 [110], Inception v3 [111], and YOLOS [112], for the task of parasite egg detection. The experiments were conducted using the Chula-ParasiteEgg-11 dataset. Among the models tested, YOLOv8 optimized with the SGD optimizer demonstrated superior performance, achieving a mAP of 0.92 and an F1-score of 98% [113].

One more research leveraged the Chula-ParasiteEgg dataset to train and evaluate various models, including CNNs and CoAtNet. Initial experiments explored common CNN architectures such as EfficientNet and DenseNet with varying layer depths. Subsequently, the study investigated the potential improvements offered by vision transformers through a self-attention mechanism. The final focus was on demonstrating the superior performance of CoAtNet [72], which integrates attention and convolution techniques in a single framework, for the classification of parasitic eggs. The results highlighted the high recognition capability of the proposed CoAtNet model, which achieved an average accuracy of 93% and an average F1 score of 93% [114].

A large comparison experiment, which involved extensive image pre-processing and augmentation to enhance the dataset, utilized variants of YOLOv5, due to its faster speed and lower memory usage, for detection and classification. The authors managed to collect a dataset of 5393 images of five different species of parasites to conduct their tests on. The performance of the system was evaluated based on various parameters. Notably, the algorithms achieved an impressive mAP of approximately 97% and a rapid detection time of 8.5 milliseconds per sample. The researchers then went on to compare these results with four other common object-detection models, namely SSD, Faster-RCNN, AlexNet and ResNet50. Notably, the YOLOv5 model outperformed the rest, on average. [115].

Finally, a study focused on enhancing the detection of Enterobius vermicularis eggs through a

CNN, benchmarked against several leading architectures. A dataset of 2000 images originally, augmented to 40000 total digitized images, comprising E. vermicularis eggs (class 1) and artifacts (class 0), was created and augmented for training. Artifacts refer to non-egg particles or debris that could be mistaken for eggs but are irrelevant to the diagnostic task. The dataset was split using an 80:20 training–validation ratio and evaluated via five-fold cross-validation. In addition to the in-house developed CNN, the following models were evaluated for comparative performance: Xception [116]: Known for its depthwise separable convolutions, offering efficiency in processing complex images. MobileNet [117]: Lightweight and efficient, designed for use in resource-limited environments. EfficientNetB1: Balances complexity and accuracy through its scalable architecture. DenseNet121: Uses a unique connectivity pattern to enhance gradient flow and improve performance. InceptionV3: Employs hybrid convolutional architecture with multiple kernel sizes to capture multi-scale features. ResNet50: Utilizes a deep residual learning framework, enabling the effective training of deeper networks. The proposed CNN demonstrated notable improvements after data augmentation, achieving 90.0% accuracy, precision, recall, and F1-score, with its ROC-AUC increasing from 0.77 to 0.97, indicating improved stability and reliability. Despite its smaller file size, the CNN model produced respectable results when comparing it to larger architectures. Among the models, Xception, with an almost 100 times larger file size, achieved the highest performance, with 99.0% accuracy, precision, recall, and F1-score, showcasing its superior capability for detecting E. vermicularis eggs [118].

3.2.2 Studies on Canine and/or Feline Parasites

An early study which acknowledged the research gap in veterinary parasitology, focused on canine intestinal parasites and had the goal of detecting four genera of parasites infecting dogs, specifically *Toxocara spp.*, *Ancylostoma spp.*, *Trichuris spp.*, and *Giardia spp.*. Their image processing algorithm was divided into three main steps: image segmentation, object representation, and object recognition. During image Segmentation, an enhanced image was computed to increase contrast, making parasites and similar impurities appear brighter while darkening the rest of the image. The enhanced image was then thresholded, and objects with areas outside the typical parasite range were eliminated. The remaining objects were submitted for feature extraction. For object representation, the features were extracted based on color, shape, and texture, providing a descriptive representation of each segmented object. Finally, for object recognition, a Support Vector Machine classifier was trained to differentiate between impurities and each of the four species of canine intestinal parasites. To train the SVM classifier, a dataset of 10,699 instances was utilized, comprising 3,132 parasite instances and 7,567 fecal impurities. The impurities primarily consisted of cells, organic matter, and plant structures from undigested digestive waste. The researchers reported among others a 52.88% average positive prediction rate with a Cohen's Kappa value of 0.7636 for their new system [119].

A following study which also had its focus on dogs, aimed to detect and identify roundworm and whipworm eggs in fecal samples using image processing techniques. The authors chose to employ a CNN with the AlexNet architecture for training. The system's performance was evaluated based on a

confusion matrix. A total of 410 images were analyzed, with the model achieving an accuracy of 96% for whipworm egg classification and 90.74% for roundworm egg classification [120].

A noteworthy study presented a microscopy image analysis pipeline for gastrointestinal parasites in cats and dogs, utilizing deep learning methods. In dogs, the parasites were: Ancylostoma spp., Cystoisospora spp., Giardia spp., Toxocara spp., and Trichuris spp.. For cats, they were: Ancylostoma spp., Cystoisospora spp., and Platynosomum spp.. Approximately 1,600 images of these parasites were used for the study. The authors proposed a pipeline for detecting and identifying the parasites present in their images. The proposed pipeline consists of three sequential steps. Firstly, the detection step identifies objects within images that are likely to be parasitic structures, using the method of U^2Net [121]. Then, the component isolation step extracts detected objects and isolates them into artificially created images. Finally, the component identification step classifies each isolated component by species, with impurities included as a possible class. To classify each component, the model was fine-tuned from a pre-trained ImageNet model, adapted to an in-house dataset. This dataset was comprised of 1,791 segmented images of human parasites and 6,068 images of impurities. Several deep learning networks were tested, including Vgg-16 [62], DenseNet-121 [122], DenseNet-169 [122] and MobileNetV2 [65]. The pipeline demonstrated a combined detection and identification accuracy of 95% for the cat dataset and 84.3% for the dog dataset [123]. Notably, the authors mentioned that the most challenging species for both detection and identification was the Giardia spp..

Lastly, a research group assessed the performance of a commercially available system, consisting of a sample preparation device, a scanner and an analysis software, able to automatically detect and identify parasites from fecal samples of dogs and cats. The system was tested on 2191 fecal samples of cats and dogs, which included eggs, cysts and oocysts of the species *Ancylostoma*, *Toxocara*, *Trichuris*, *Cystoisospora* and *Giardia*, with two different scanner models. The authors did not clearly specify the details of the deep learning algorithm utilized, although it was implied that it is utilizing some variant of the YOLO family of models, perhaps YOLOv3 or later. It was reported that the performance of the algorithm closely matched that of the parasitologists, with sensitivities in the range of 80.0–97.0% and specificities reaching the range of 93.7–100.0%. Additionally, the researchers regarded the Giardia species as one of the more difficult gastrointestinal parasites to identify through fecal examination due to its small size and translucent nature [124].

3.2.3 Positioning of the Present Thesis

While previous studies have made significant contributions toward the automation of intestinal parasite detection in canine and feline fecal samples, they often rely on either handcrafted features or traditional classification architectures with limited generalizability. For instance, some earlier works employed SVMs on manually extracted features, which, while effective in constrained settings, lack the robustness and scalability of deep learning-based approaches. Others, such as those utilizing AlexNet or transfer learning on cropped objects, focused primarily on classification tasks post-segmentation and often evaluated performance on relatively static datasets.

In contrast, the present study contributes to the field by proposing an end-to-end object detection framework that eliminates the need for manual feature extraction or prior component segmentation. In addition, utilizing a modern, lightweight YOLO architecture, this implementation enables real-time detection and classification directly from full microscopy images—an important distinction that better reflects practical, clinical conditions. In doing so, this work places particular emphasis on deployment-readiness and considers hardware deployment and usability.

Moreover, this work is one of the very few that focuses on veterinary parasitology, making use of a curated dataset comprised of key species of canine and feline parasites with serious importance to pet and overall public health.

Finally, the integration of model explainability, an aspect often underexplored in the domain of parasite detection is another distinction of this study. Examples of the model's decisions are made visually transparent highlight the specific regions within microscopy images that contributed most significantly to a given prediction, ensuring that model behavior can be qualitatively evaluated against known morphological features of each parasite species. This approach aligns with the increasing emphasis in medical AI on human-in-the-loop systems, where algorithmic assistance complements but does not replace expert judgment.

4 Idea and Approach

4.1 Motivation

Parasitic infections remain a significant global health burden, particularly in low-resource settings where access to timely and accurate diagnostic services is limited. Microscopic examination of fecal samples or blood smears is considered a gold standard for detecting many parasitic organisms. However, this manual process is labor-intensive, time-consuming, and highly dependent on the experience and skill of the microscopist. As a result, diagnostic accuracy can vary widely across individuals and institutions, contributing to both false negatives and false positives.

Parasitic infections in domestic animals, particularly canines and felines, pose significant health challenges not only to the animals themselves but also to humans, due to the zoonotic nature of some of them. Infections caused by *Toxocara spp.*, *Giardia spp.*, *Hookworms* and other intestinal parasites are common in household pets and are frequently transmitted through contaminated environments, food, or direct contact with infected animals. These common parasites pose an added burden due to their ability to infect humans, leading to conditions such as visceral and ocular larva migrans or chronic gastrointestinal issues. In this context, misdiagnosis or underdiagnosis in animals can have direct public health consequences, especially in urban and peri-urban settings where human–animal interactions are frequent.

By focusing on a dataset comprising microscopic images of canine and feline parasites, the proposed system directly addresses the diagnostic gap in veterinary parasitology. This work is particularly relevant in urbanized regions where pet ownership is high and close proximity between animals and humans increases the risk of zoonotic transmission. It also contributes to public health by helping to interrupt transmission cycles through timely diagnosis and treatment in animal hosts. Furthermore, the insights and model architectures developed in this research can be adapted for future applications in other areas of veterinary diagnostics, wildlife monitoring, or even human parasitology, making it a foundational contribution to digital pathology in infectious disease control.

4.2 Problem Statement

The challenge of automating parasite detection through computer vision lies in the inherent complexity and variability of microscopic images. Parasites often appear in different shapes, sizes, and orientations depending on their life stages, species, and staining techniques. Moreover, background noise, overlapping structures, and artifacts in slides introduce further ambiguity, making the task non-trivial even for trained professionals.

Traditional image processing techniques, while effective in controlled environments, often fail to generalize across different datasets and staining protocols. In recent years, the integration of artificial intelligence and specifically, computer vision into medical diagnostics offers a promising path toward enhancing diagnostic efficiency and accuracy. Computer vision systems have demonstrated remarkable

performance in tasks such as image classification, object detection, and semantic segmentation, with applications now spanning radiology, dermatology, ophthalmology, and histopathology. However, their application to microscopic parasite detection remains relatively underexplored largely due to the niche nature of the domain and presents unique challenges, such as:

- The need for high precision in detecting complex visual data such as small, densely packed targets within cluttered visual fields.
- Limited availability of large, well-annotated datasets specific to parasitic infections.
- The requirement for lightweight and real-time models suitable for deployment in clinical or field environments with constrained computational resources.

Given these challenges, this study aims to present a deep learning-based solution that can robustly detect and localize parasitic organisms in microscopic images. Specifically, the goal is to fine-tune an object detection model on a custom dataset of parasitic images. The rationale for choosing the appropriate model is its balance between detection accuracy and real-time performance, which is crucial for practical deployment in diagnostic workflows. The proposed approach seeks to enhance diagnostic accuracy, assist veterinarians in rapid screening of samples, reduce inter-observer variability and overall contribute toward improving the early detection of parasites with zoonotic potential by integrating artificial intelligence into point-of-care medical diagnostics.

The problem can be expressed as follows. Let a set of microscopy images be denoted by:

$$\mathcal{I} = \{I_1, I_2, \ldots, I_N\}, \quad I_i \in \mathbb{R}^{H \times W \times 3}$$

Each image may contain zero or more parasites of different species, each annotated with a bounding box and class label. Define the ground truth annotations for image I_i as:

$$\mathcal{G}_i = \{(b_{i1}, c_{i1}), (b_{i2}, c_{i2}), \ldots, (b_{in_i}, c_{in_i})\}$$

where:

- $b_{ij} \in [0, 1]^4$ represents a normalized bounding box (x, y, w, h)
- $c_{ij} \in \{1, 2, 3, 4\}$ is the class label (e.g., *Giardia*, *Hookworm*, etc.)
- C is the total number of parasite classes

The goal is to learn a detection function f_θ, parameterized by deep neural network weights θ, such that:

$$f_\theta(I_i) \rightarrow \hat{\mathcal{G}}_i = \{(\hat{b}_{i1}, \hat{c}_{i1}, s_{i1}), \ldots\}$$

where:

- $\hat{b}_{ij}$ is the predicted bounding box
- $\hat{c}_{ij}$ is the predicted class
- $s_{ij} \in [0, 1]$ is the confidence score for the detection

4.3 Dataset

The dataset used in this thesis was collected in parallel with its conduction, within the Laboratory of Parasitology and Parasitic Diseases, School of Veterinary Medicine, Faculty of Health Sciences, Aristotle University of Thessaloniki. The dataset consists of 310 RGB images distributed across four classes of canine and feline species of parasites: Toxocara, Cystoisospora, Hookworm, Giardia. The images were captured either using a smartphone camera pointing through a microscope lens, or with the use of a dedicated microscope camera with appropriate software. All images were taken through a lens with a magnification factor of x40 and have a resolution of either 3000×4000 or 2560×1920 pixels. The images were captured either in JPG or TIF format, with the latter then being converted to JPG in order to match the model input requirements. The per-class distribution of images as well as instance counts are presented in Table 1.

Bounding box annotations are employed in this study to label and localize parasites within microscopic images. These boxes are carefully drawn to tightly enclose each visible parasite, providing spatial information necessary for object detection models to learn the position and scale of the target entities. Each bounding box is also assigned a class label corresponding to the specific type of parasite, enabling both localization and classification tasks to be carried out concurrently. The annotation process was applied to all images containing at least one partially visible parasite to ensure the dataset was comprehensive and informative. Images were carefully annotated by the author under the guidance and approval of an expert in veterinary parasitology. To ensure consistency, all images containing ambiguous samples were reviewed in detail and sometimes excluded from the final dataset.

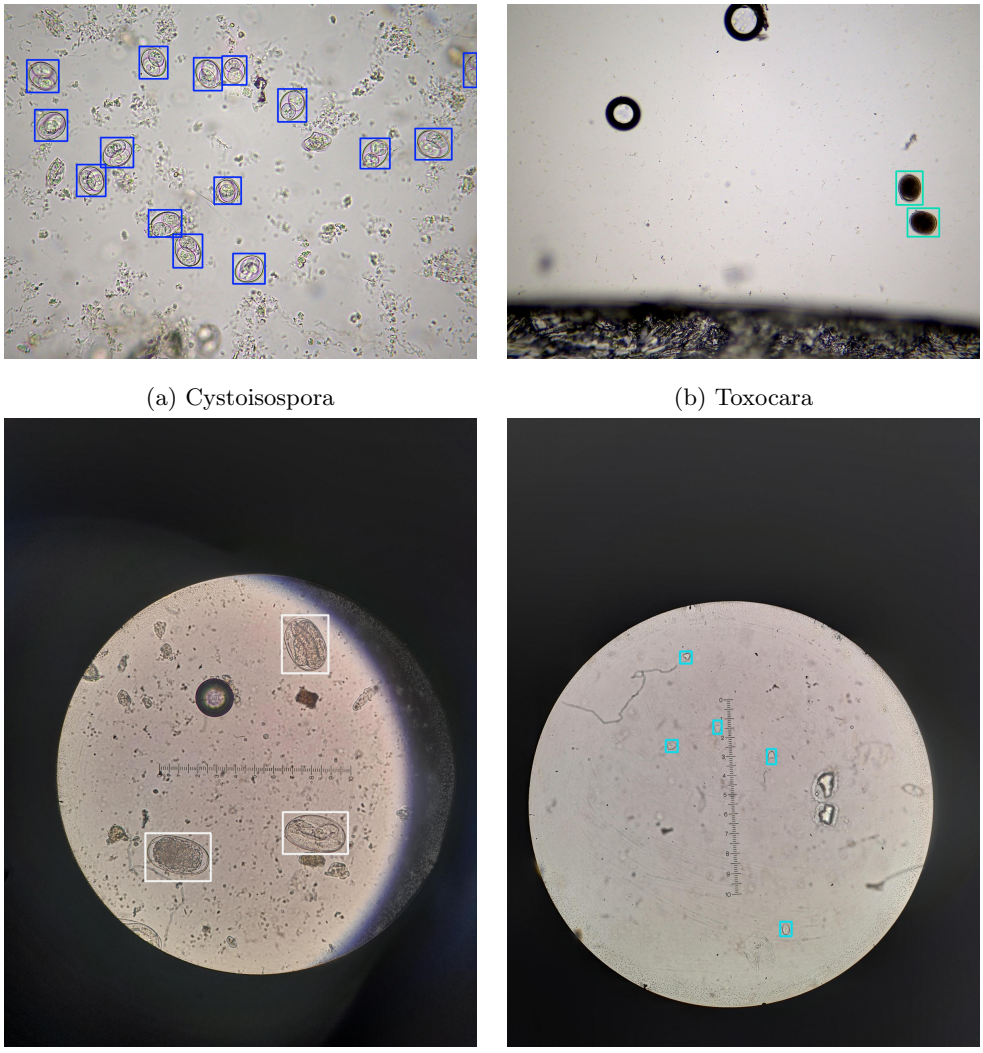
Table 1: Dataset Distribution

Classes	**Images**	**Instances**
Toxocara	162	248
Cystoisospora	58	315
Hookworm	64	72
Giardia	26	361
Total	**310**	**996**

The dataset is specifically curated for the detection of parasitic organisms in canine and feline fecal samples. Each image captures detailed visual information essential for identifying various types of parasites, making the dataset a valuable resource for advancing automated diagnostic techniques.

The primary objective behind assembling this dataset is to support the development and validation of AI-assisted diagnostic models that can accurately detect and localize parasite structures in digital microscopy images.

By focusing on canine and feline samples, the dataset addresses a significant need in veterinary parasitology, where rapid and accurate detection is crucial for effective treatment and disease management. The annotated images provide the necessary ground truth for training object detection algorithms, such as YOLO-based models, enabling them to learn the morphological characteristics of different parasites. In doing so, the dataset facilitates the creation of robust, generalizable diagnostic tools that can be deployed in clinical or field settings, enhancing the efficiency and consistency of parasitological assessments.



(a) Cystoisospora

(b) Toxocara

(c) Hookworms

(d) Giardia

Figure 11: Examples of bounding box annotations.

The images presented in Figure 11 illustrate representative examples of bounding box annotations used during dataset preparation for object detection. Each image corresponds to samples of a distinct

parasite class: Cystoisospora, Toxocara, Hookworm, and Giardia. The bounding boxes, delineated in consistent color-coded schemes, tightly enclose parasite eggs or cysts, enabling the model to learn precise spatial features. Notably, the annotations vary in scale, contrast, and background complexity, reflecting realistic conditions under which these organisms are observed microscopically. This variability in the annotated dataset is crucial for enhancing the model's generalization capacity and robustness when deployed in practical diagnostic scenarios.

By quantifying the number of annotated objects per image, the object density of the dataset is presented (Figure 12). The distribution is heavily right-skewed, with the majority of images containing a small number of objects. Specifically, the most frequent count is a single object per image (187 images), while 52 images contain two objects. A smaller subset of the dataset features higher-density scenes with more than 10 objects per image, which are particularly valuable for evaluating detection performance under challenging conditions such as occlusion, overlap, and visual clutter. Quantitatively, the distribution exhibits a mean of 3.58 objects per image, a median of 2, and a standard deviation of 5.17, indicating high dispersion. The minimum and maximum number of objects per image are 1 and 43, respectively. This variability highlights the heterogeneous nature of object density across the dataset and has important implications for model generalization, especially in real-world deployments where image complexity may differ significantly from the training set. In such contexts, models trained on datasets with a wide range of densities are more likely to exhibit robust performance.

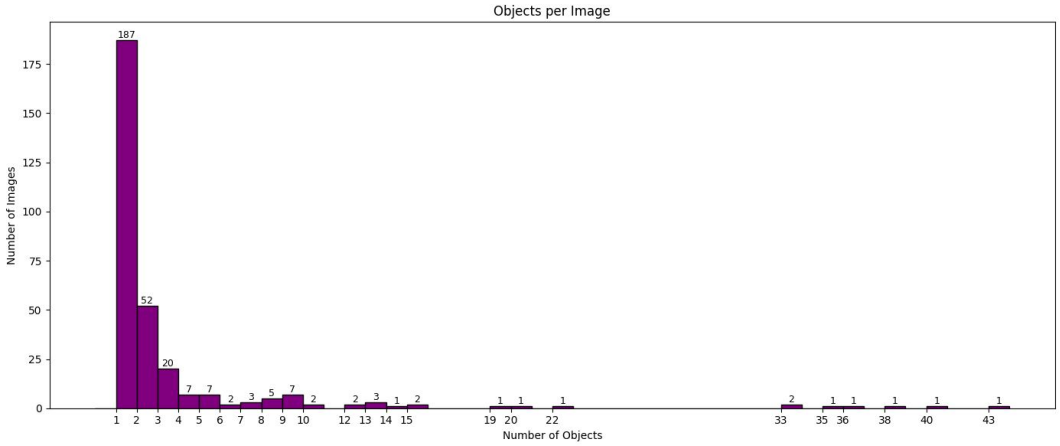


Figure 12: Distribution of annotated object counts per image.

Figure 13 presents a normalized heatmap of object center locations aggregated across the dataset. The distribution is moderately uniform with a slight concentration toward the central region of the image plane (normalized coordinates $x, y \in [0, 1]$). This central bias is expected in many visual datasets, often reflecting acquisition or annotation tendencies where salient objects are captured near the center of the field of view.

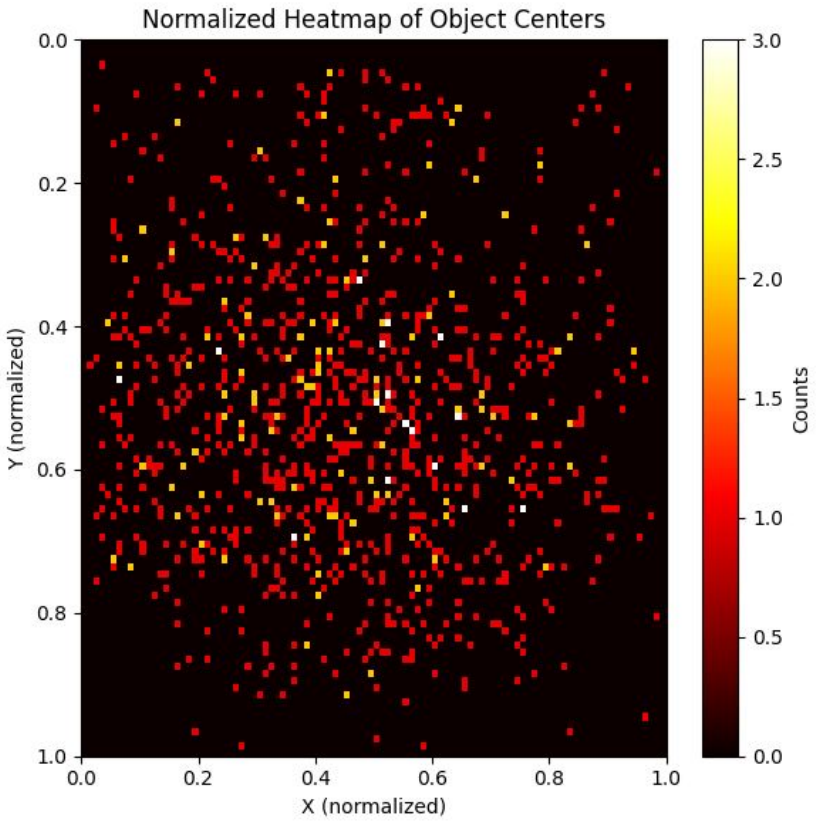


Figure 13: Normalized 2D heatmap of object center locations across all annotations. Brighter regions indicate higher density.

Importantly, the heatmap also demonstrates a substantial spread of object centers across the spatial extent of the image, indicating that the dataset offers reasonable spatial diversity. This characteristic is beneficial for training object detectors that generalize well to objects at varying locations and scales. Moreover, the absence of strong corner clustering suggests minimal positional bias, which is advantageous for robust learning in anchor-free detection architectures.

4.4 Methodology

This section outlines the methodology adopted in the present thesis, focusing on the development and application of a computer vision model tailored for the field of parasitology. Automating the detection and annotation of parasites in real-time microscopy holds significant potential as a decision-support tool for veterinary professionals. In clinical and diagnostic settings, the ability to rapidly and accurately identify parasitic organisms within fecal samples is essential for timely treatment and disease management. The central objective is to develop and evaluate the effectiveness of an automated system for analyzing microscopic images, particularly in the context of parasite diagnosis. By leveraging advancements in deep learning and object detection techniques, the methodology aims to enhance diagnostic accuracy, reduce the burden on experts, and facilitate scalable, real-time analysis in clinical

or research settings.

4.4.1 Problem Formulation

The detection of parasitic structures within microscopic samples is framed as an object detection task. This approach enables a model to not only identify the presence of parasitic organisms but to also accurately localize them within the images. To address it, the YOLOv11 algorithm was employed to develop a robust object detection model tailored to parasitological diagnostics. Initially, the dataset underwent annotation using bounding boxes to localize regions of interest, as outlined in Section 4.3. Each bounding box was defined by the coordinates of its top-left corner $(x_{\min}, y_{\min})$ and its bottom-right corner $(x_{\max}, y_{\max})$. These coordinates provided the spatial extent of each detected parasite within the image, serving as essential ground truth for subsequent training and evaluation phases in the object detection pipeline.

In order for a model to learn a given detection function, the parameters of this function, the weights of the model, need to be accurately configured. This is addressed by training the model with the goal of minimizing an appropriate loss function through backpropagation. More precisely, the YOLOv11 model is trained using a composite loss function that balances multiple objectives critical for accurate object detection. This loss function $\mathcal{L}$ is be defined as:

$$\mathcal{L} = \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}+\lambda_{\text{dfl}}}\mathcal{L}_{\text{dfl}}$$

The weighting coefficients λ_{box}, λ_{cls} and λ_{dfl} are hyperparameters that control the relative influence of each loss term, allowing the model to prioritize localization, classification, or box coordinate regression as needed.

Box Loss $\mathcal{L}_{\text{box}}$ represents the bounding box regression loss, which ensures precise localization by minimizing the error between predicted and ground truth bounding boxes:

$$\mathcal{L}_{\text{box}} = 1 - \text{CIoU}$$

- **Complete Intersection over Union (CIoU) Loss**: This loss measures the overlap between the predicted and ground truth boxes, considering factors like center distance and aspect ratio, ensuring better geometric alignment. Let:
  - $\mathbf{b}_{\text{pred}} = (x, y, w, h)$: Predicted bounding box (center coordinates, width, height)
  - $\mathbf{b}_{\text{gt}} = (x_{\text{gt}}, y_{\text{gt}}, w_{\text{gt}}, h_{\text{gt}})$: Ground truth bounding box
  - IoU: Intersection-over-Union
  - c^2: Squared diagonal length of the smallest enclosing box covering both predicted and ground truth boxes
  - ρ^2: Squared Euclidean distance between the centers of predicted and ground truth boxes

Then, the **CIoU** loss is defined as:

$$\text{CIoU} = \text{IoU} - \frac{\rho^2(\mathbf{b}_{\text{pred}}, \mathbf{b}_{\text{gt}})}{c^2} - \alpha v$$

Where:

$$v = \frac{4}{\pi^2}\left(\arctan\left(\frac{w_{\text{gt}}}{h_{\text{gt}}}\right) - \arctan\left(\frac{w}{h}\right)\right)^2$$

$$\alpha = \frac{v}{1 - \text{IoU} + v}$$

The αv term adjusts the loss to account for differences in aspect ratios, improving convergence and box alignment.

Distribution Focal Loss (DFL) $\mathcal{L}_{\text{dfl}}$ treats bounding box sides (left, top, right, bottom) as discrete probability distributions over bins instead of predicting continuous values directly. This allows for better localization accuracy and quantized regression. Let:

- $x \in \mathbb{R}$: Ground truth continuous value (e.g., for a box side).
- $p \in \mathbb{R}^M$: Predicted probability distribution over M discrete bins (e.g., $M = 16$ or 64).
- $\lfloor x \rfloor = l$, $r = l + 1$: Nearest lower and upper bin indices.
- $w_l = r - x$, $w_r = x - l$: Linear interpolation weights.

Then:

$$\text{DFL}(x, p) = w_l \cdot \text{CE}(p, l) + w_r \cdot \text{CE}(p, r)$$

Where:

$$\text{CE}(p, i) = -\log(p_i)$$

is the cross-entropy loss for bin i.

Classification Loss $\mathcal{L}_{\text{cls}}$ represents the classification loss, which penalizes incorrect class predictions and guides the model toward better discrimination among object categories. Binary Cross-Entropy (BCE) with Logits Loss is a single numerically stable operation which applies BCE to raw, unbounded network outputs (logits) and is typically combined with the sigmoid activation function. This classification function also allows for multi-label classification, where each sample can belong to multiple classes simultaneously. Let $x \in \mathbb{R}$ denote the raw prediction (logit), and $y \in \{0, 1\}$ be the binary ground truth label. The sigmoid function is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}},$$

which maps logits to the open interval $(0, 1)$, interpreting them as probabilities. The binary cross-entropy loss is then given by:

$$\mathcal{L}_{\text{BCE}}(x, y) = -\left[y \log(\sigma(x)) + (1 - y) \log(1 - \sigma(x))\right].$$

The model was then trained on the annotated microscopy images, enabling it to autonomously draw bounding boxes around different species of parasites and accurately classify them.

4.5 Evaluation Metrics

In object detection tasks, evaluating model performance requires more than assessing simple accuracy. Unlike classification problems, object detection involves both localizing objects within an image and correctly identifying their classes, making evaluation inherently more complex. Metrics in this domain are designed to capture the dual aspects of spatial precision and categorical correctness, accounting for factors such as overlapping predictions, missed detections, and false positives. These metrics provide a comprehensive framework to quantify how well a model performs in realistic, often noisy, visual environments.

Intersection over Union is a fundamental object detection metric employed to assess the degree of overlap between a predicted bounding box and the corresponding ground truth bounding box. It is calculated as the ratio of the area of intersection to the area of union between the predicted and ground truth boxes, serving as a critical evaluation criterion in object detection tasks. It is calculated as:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

where the Area of Overlap is the region where the predicted box and the ground truth box intersect, and the Area of Union is the total area covered by both boxes.

Detection outcomes need to be clearly defined in the context of object detection before moving on to more complex metrics. These are the true positives (TP), false positives (FP) and false negatives (FN) of object detection.

- A **true positive** in object detection occurs when a predicted bounding box correctly localizes and correctly classifies an object. Let B_p be a predicted bounding box and B_g a ground-truth box. Then, B_p is a true positive if:
  - The predicted class matches the ground-truth class.

  - The IoU between B_p and B_g exceeds a predefined threshold τ (commonly $\tau = 0.5$):

$$\text{IoU}(B_p, B_g) = \frac{\text{area}(B_p \cap B_g)}{\text{area}(B_p \cup B_g)} > \tau$$

  - Each ground-truth box can be matched to only one predicted box.

- A **false positive** occurs when a predicted bounding box does not correctly correspond to any ground-truth object:
  - The IoU is below the threshold.
  - The class prediction is incorrect.
  - The prediction is a duplicate detection of the same object.
- A **false negative** occurs when a ground-truth object is missed by the detector.
  - No predicted box satisfies the IoU threshold with the correct class label of a given object.

Precision is defined as the proportion of true positive predictions to the total number of positive predictions, including both true positives and false positives. It is particularly critical in scenarios where the cost associated with false positives is high.:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (also referred to as **Sensitivity**) quantifies the proportion of true positive predictions relative to the total number of actual positive instances, encompassing both true positives and false negatives. Recall is especially important in contexts where the consequences of false negatives are substantial:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score represents the harmonic mean of precision and recall, offering a single comprehensive metric that balances the trade-off between these two aspects. It is particularly valuable in scenarios involving imbalanced class distributions, where optimizing both precision and recall is critical:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Average Precision is the area under the Precision–Recall (PR) curve, a fundamental metric for evaluating object detection performance per class.

Let precision(r) denote the precision at recall r. Then:

$$\text{AP} = \int_0^1 \text{precision}(r)\, dr$$

Mean Average Precision (mAP) is a comprehensive evaluation metric that computes the mean of the Average Precision scores across all classes. It offers an overall assessment of a model's performance, particularly in complex tasks such as object detection, where accurate multi-class predictions are essential:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i$$

where N is the number of classes, and AP_i is the Average Precision for class i.

Mean Average Precision at IoU thresholds (mAP@50–95) measures the mean of the Average Precision scores calculated across multiple IoU thresholds for classifying a detection as positive, ranging from 0.50 to 0.95 in increments of 0.05. This metric offers a more rigorous and comprehensive evaluation of a model's detection performance by assessing its accuracy across varying degrees of detection precision:

$$\text{mAP@50-95} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i$$

where N is the number of classes, and AP_i is the Average Precision for class i, calculated at multiple IoU thresholds (from 0.50 to 0.95).

5 Implementation and Results

In this chapter, the technical details and experimental procedures carried out in this study are presented. The implementation of the proposed training configurations is discussed in depth. Furthermore, the performance of the models is evaluated through a comprehensive set of experiments on training, validation, and test datasets. Metrics such as precision, recall, mAP, and loss curves are used to analyze the effectiveness and generalization capabilities of the system. Comparative results across various model configurations are also included to support the conclusions drawn from the experimental observations.

5.1 Experimental Setup

All training experiments were conducted on the Google Colab platform, utilizing a single NVIDIA T4 GPU with 15GB of VRAM. The YOLO implementation was adapted from the Ultralytics codebase with modifications made to support fine-tuning on the custom parasite dataset.

5.2 Training Pipeline

5.2.1 Data Preparation

The annotated dataset was divided into training, validation, and test subsets using an approximate 70-20-10 split, based on object instances (Table 2). This division was designed to support robust model development while ensuring a fair evaluation of performance. Special attention was given to maintaining class balance across the splits to the greatest extent possible, as imbalanced datasets can bias the model toward overrepresenting certain classes while underperforming on others. By preserving representative distributions of parasite types in each subset, the model was better positioned to generalize well to new, unseen data, and the validation and test results provided a more accurate reflection of real-world performance.

Classes	**Instances**	**Train**	**Validation**	**Test**
Toxocara	248	174	48	26
Cystoisospora	315	223	60	32
Hookworm	72	50	15	7
Giardia	361	247	74	40
Total	**996**	**694**	**197**	**105**

Table 2: Dataset Split Instances

Due to computation restrictions, input images were downsized by the model to 1440x1440, during training. Different training setups with resize resolutions that retained the original aspect ratio of images were experimented with but were found to produce inferior results. Lower resolutions were

also experimented with and indicated that the higher input resolution - the higher the performance, at the cost of speed. The final resolution was determined to be a satisfactory balance between accuracy, training times and resource utilization. During inference, a similar resizing takes place, with the notable difference that the images retain their original aspect ratio, by resizing their largest dimension to 1440 pixels and their smallest in an accommodating way such that the final aspect ratio is the same.

5.2.2 Data augmentation

Data augmentation is a fundamental strategy in modern computer vision pipelines, particularly in tasks involving limited or imbalanced datasets. By synthetically expanding the training data distribution through label-preserving transformations, augmentation introduces variability that improves the model's ability to generalize to unseen data. It helps mitigate overfitting by exposing the network to diverse appearances of the same class, thereby encouraging the learning of robust, invariant features rather than memorizing specific instances.

In parasite microscopy, data augmentation plays a critical role in enhancing model performance, particularly due to the limited availability of annotated datasets and the high variability in sample appearance. Microscopy images often suffer from challenges such as uneven staining, varying magnifications, background artifacts, and morphological diversity among parasite species and life stages. Data augmentation mitigates these issues by synthetically expanding the training set through biologically plausible transformations that preserve class semantics. Techniques like flipping, translation, and color jitter simulate variations in sample preparation, slide orientation, and lighting conditions commonly encountered in laboratory workflows. More advanced augmentations such as mosaic composition introduce contextual and compositional diversity, helping the model learn robust features despite the sparsity or irregularity of parasite occurrences. This is particularly valuable in tasks involving dense object detection or segmentation of small, morphologically subtle targets. By enriching the training distribution, augmentation reduces overfitting, improves generalization to unseen samples, and enhances the model's resilience to noise and domain shifts, which are frequent in real-world parasitology workflows.

In detail, the set of data augmentation techniques that yielded the best results were:

- **HSV Augmentation:** Saturation, which modifies the intensity of colors in the image and brightness value, which changes the brightness of the image, were factored by up to ± 0.5 and ± 0.1 respectively, while color hue remained fixed at 0.0.
- **Translation:** Translation, which shifts images horizontally and vertically by a random fraction of the image size, helps models learn to detect partially visible objects and improves robustness to object position. Images were randomly translated by up to 10% of their dimensions.
- **Flipping:** Flipping the image horizontally, along the x-axis, or vertically, along the y-axis, helps the model generalize to objects appearing in mirrored or upside-down orientations, respectively. Both horizontal and vertical flips were applied with a probability of 0.5.

- **Mosaic:** Full mosaic augmentation was enabled (= 1.0), combining four images into one to increase diversity and scale robustness. This advanced augmentation technique is highly effective for improving small object detection and context understanding.
- **Disabled Augmentations:** Rotation, shear, scaling, and perspective transformations were disabled, as they are not representative of real-world scenarios of microscopy tasks. MixUp, CutMix, and random erasing were also not used.

5.2.3 Model Configuration

The model was initialized with weights pretrained on the COCO dataset [125] for fine-tuning. The fine-tuning stage was dedicated to adapting the YOLOv11-nano model to the specific task of detecting parasites in microscopic images. This process involved training the model on the annotated images, which were downsized by the model to 1440x1440 pixels. These images were fed into the YOLOv11 architecture, enabling the model to learn to distinguish parasitic forms from surrounding structures and artifacts commonly found in fecal microscopy.

5.3 Evaluation & Results

Model performance was evaluated using standard object detection metrics such as Precision, Recall, F1-score, mAP@0.5, mAP@[0.5-0.95]. Loss curves were recorded during training to monitor convergence and overfitting. Additionally, per-class metrics and confusion matrices were computed.

Validation To ensure optimal model performance, an extensive hyperparameter tuning process was undertaken. This involved systematic experimentation with both model hyperparameters and regularization techniques. Key hyperparameters such as batch size and learning rate were rigorously explored. Batch sizes of 8, 16, 32, and 64 were tested, along with initial learning rates ranging from 0.0001 to 0.1 and final learning rates from 0.00001 to 0.01. Several optimization algorithms were evaluated, including Adam, AdamW, and Stochastic Gradient Descent, with corresponding adjustments to weight decay values where applicable.

Additionally, the impact of learning rate scheduling was investigated, although it did not result in measurable improvements in validation performance. For each training configuration, key metrics—namely training and validation loss, mAP@50 and mAP@50–95 were recorded across all epochs. Training was conducted for a maximum of 100 epochs, with early stopping employed to prevent overfitting, using a patience threshold of 20 epochs based on validation mAP@50–95 following the COCO evaluation protocol.

The process ultimately resulted in an optimal configuration: the AdamW optimizer, a momentum coefficient of 0.9, an initial learning rate of 0.001 decaying to 0.00001, weight decay set to 0.0005, and a batch size of 16. Additionally, a warm-up phase over the first three epochs helped stabilize training, with warm-up momentum and bias learning rates set to 0.8 and 0.0, respectively.

The model evaluation on the validation set demonstrated a mAP of 96.4% at IoU=0.5 and 80.8% mAP@[0.5:0.95], indicating excellent detection performance across both easy and challenging overlap thresholds. A breakdown of per-class metrics is shown in Table 3.

Table 3: Results on Validation Split

Classes	**Precision**	**Recall**	**F1-score**	**mAP50**	**mAP50-95**
Toxocara	0.980	1	0.99	0.994	0.896
Cystoisospora	0.984	1	0.991	0.995	0.874
Hookworm	1	0.933	0.965	0.964	0.865
Giardia	0.888	0.865	0.876	0.903	0.595
Total	**0.963**	**0.95**	**0.955**	**0.964**	**0.808**

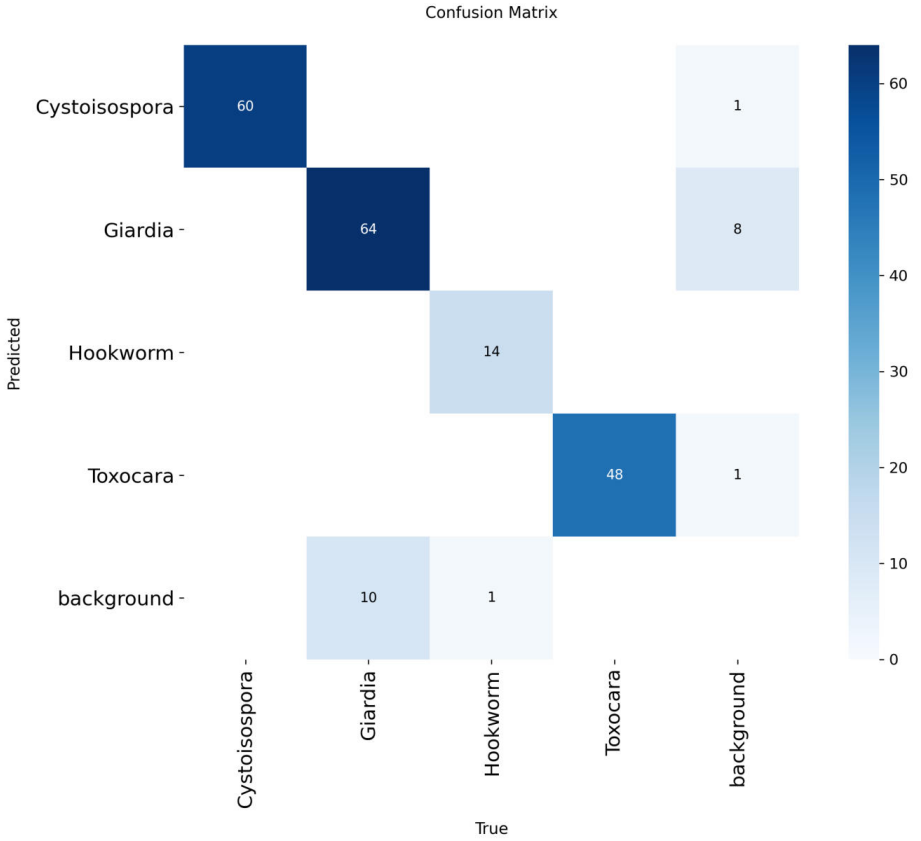


Figure 14: Validation Confusion matrix. The diagonal represents the correct (true positive) predictions. Off-diagonal predictions are either missed detections (false negative) or false predictions (false positive).

Figure 14 illustrates the validation confusion matrix. It provides some valuable early insights into model performance. The vast majority of the predictions are correct, indicated by the strong diagonal. *Giardia* seems to be the only class with moderate perfomance, with some missed detections and false

positives, but not to an alarming degree. A significant takeaway is, no class confusion is observed. Overall, the results are promising, pointing towards a successful hyperparameter selection.

The training and validation curves presented in Figure 15 illustrate the learning behavior of the model over 57 epochs. These metrics encompass the three individual loss components as well as key evaluation metrics, including precision, recall, and mAP at both IoU thresholds of 0.5 (mAP50) and 0.5–0.95 (mAP50-95).

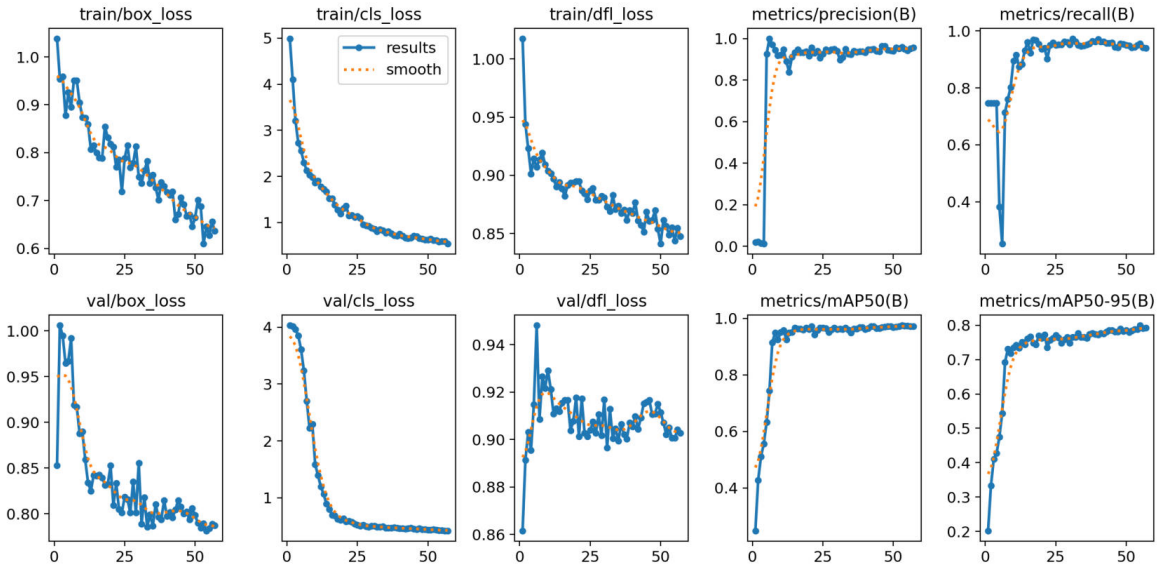


Figure 15: Train and validation curves during training. Individual loss components as well as performance metrics on the validation set are shown.

The training loss components show a consistent downward trend across epochs, suggesting that the model is successfully optimizing its parameters. Notably, all three training losses exhibit smooth convergence, with no signs of divergence or instability. The validation losses also generally decrease over time, closely following the training losses, albeit with slightly higher variance—particularly visible in the early epochs. This variance is expected due to the smaller size and more heterogeneous nature of the validation set compared to the training data. The stable behavior of the validation losses further implies that the model generalizes well to unseen data, and there is no strong evidence of overfitting.

Moreover, both mAP50 and mAP50-95 metrics demonstrate steady improvement over time, with mAP50 converging upwards of 0.9 and mAP50-95 approaching 0.80. These scores reflect excellent localization accuracy across various IoU thresholds. The gradual, monotonic increase in mAP50-95 without degradation in later epochs further suggests that the learning rate schedule and training duration were well calibrated.

Collectively, the loss and metric curves indicate a well-optimized training process with stable convergence and robust generalization. The lack of divergence between training and validation losses and the parallel trends in performance metrics suggest that the model benefits from appropriate regularization and balanced training data. Furthermore, the rapid convergence followed by a plateau

in precision, recall, and mAP suggests that the model reaches its representational capacity relatively early, implying that further performance gains may require architectural modifications or enhanced dataset diversity rather than longer training. These curves support the overall validity of the trained object detection pipeline and underscore its potential for deployment in practical diagnostic settings with unseen images.

After the selection of hyperparameters, training with validation set integration was initialized, resulting in the final model.

Testing The model was evaluated on the held-out test set and achieved a mAP of 98.9% at IoU=0.5 and 81.4% mAP@[0.5:0.95], indicating excellent detection performance across both easy and challenging overlap thresholds. The breakdown of per-class metrics is shown in Table 4. The model demonstrated high precision scores across all classes and IoU thresholds, with the exception of the *Giardia spp.* where accuracy degraded at high IoU values. This can be attributed to the fact that *Giardia spp.* are much smaller in size and is on-par with the observations of similar works [123, 124].

Table 4: Results on Test Split

Classes	**Precision**	**Recall**	**F1-score**	**mAP50**	**mAP50-95**
Toxocara	0.963	1	0.981	0.981	0.914
Cystoisospora	1	1	1	0.995	0.848
Hookworm	1	1	1	0.995	0.839
Giardia	0.95	0.95	0.95	0.985	0.653
Total	**0.978**	**0.988**	**0.983**	**0.989**	**0.814**

The Precision-Recall curve presented in Figure 16 demonstrates strong performance across all classes. Both Cystoisospora and Hookworm achieve a near-perfect mAP@0.5 score of 0.995. Giardia and Toxocara also show robust performance, with mAP@0.5 values of 0.985 and 0.981, respectively. The overall mAP@0.5 for all classes stands at 0.989, highlighting the model's excellent generalization across diverse parasite morphologies. The tight clustering of curves near the top-right corner further reflects the model's strong discriminative capacity.

The confusion matrix presented in Figure 17 provides a detailed breakdown of prediction performance across the four parasite species as well as a background class, in order to demonstrate false positive and false negative predictions. The matrix offers a more intuitive evaluation of the predicted labels on the test set, providing insight into model accuracy, class-specific precision, and inter-class confusability. The diagonal dominance of the matrix indicates true positive rates are high across the board and importantly, no class overlap is observed. Only a couple of false negatives are reported, both for the *Giardia* class. False positive rates remain low across the classes. While the quantity of samples is limited, we can reasonably conclude that the model generalizes to an acceptable degree.

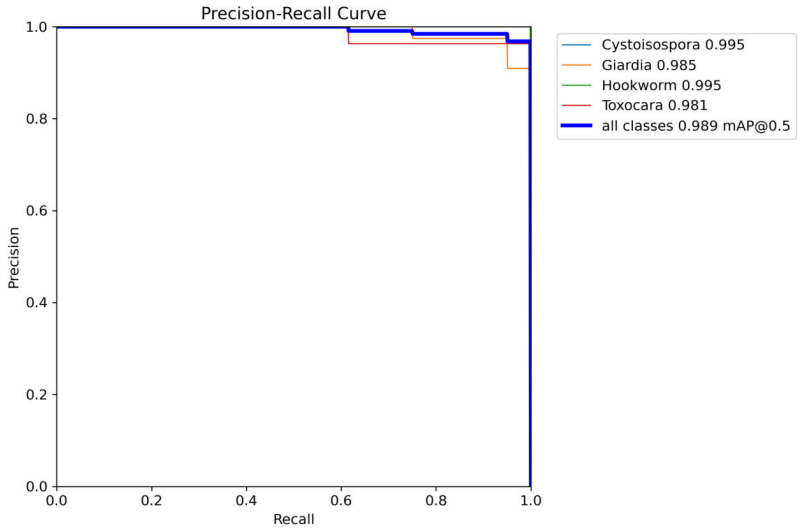


Figure 16: Precision-Recall curve. The area under the curve indicates remarkable performance.

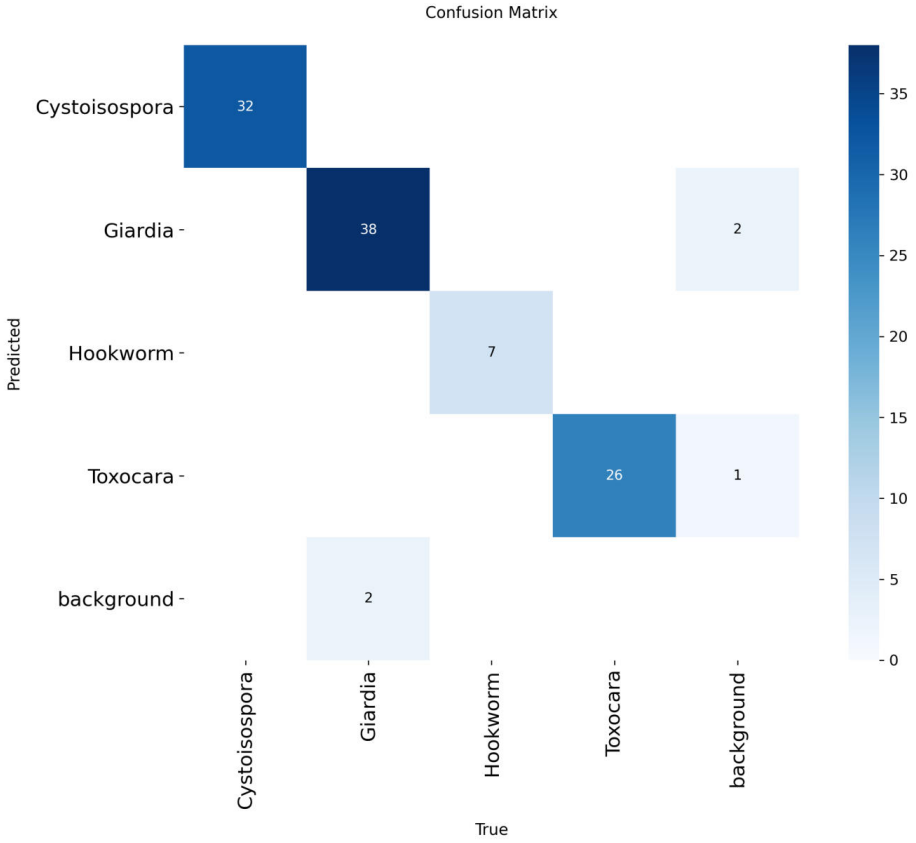
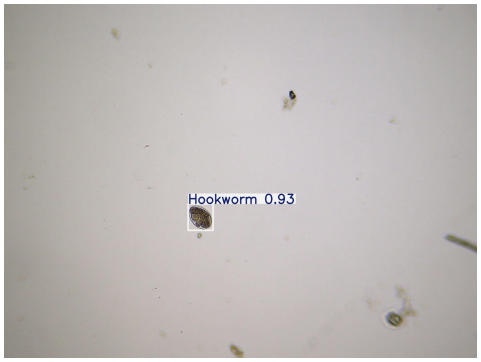
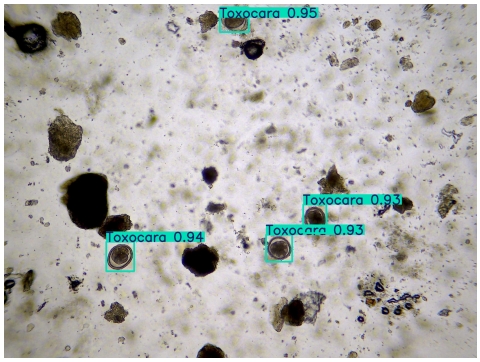


Figure 17: Test Confusion matrix. The diagonal represents the correct (true positive) predictions. Off-diagonal predictions are either missed detections (false negative) or false predictions (false positive).

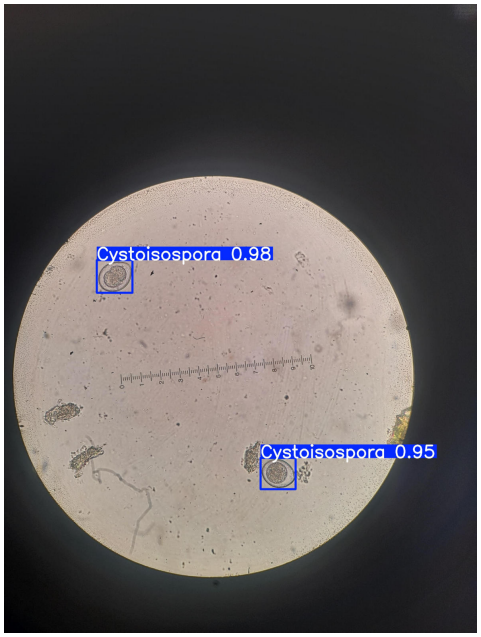
Prediction Examples Figure 18 depicts model predictions on four test set images which demonstrate successful object detection results.



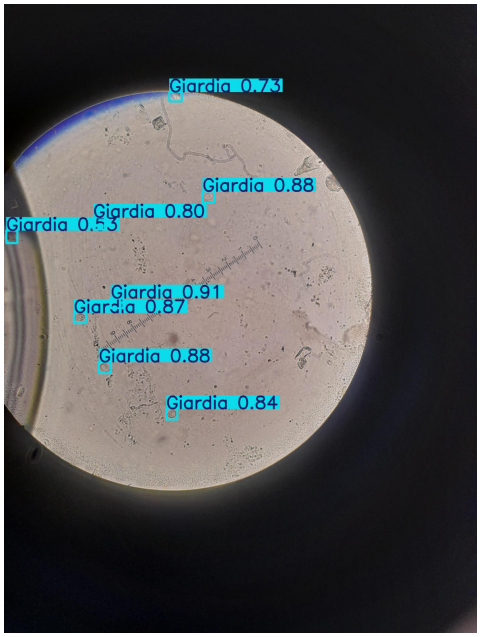
(a) A single Hookworm prediction in a clear, low-clutter background.



(b) Toxocara predictions in a noisy background. Accurate even at the edge of the image.



(c) Cystoisospora predictions. High localization accuracy despite the use of a smartphone camera.



(d) Giardia predictions in a densely populated image. Exceptional performance despite the smartphone camera and the smaller size of the samples. Displays high accuracy even around the edge of the lens.

Figure 18: Examples of correct test set predictions.

The *Hookworm* prediction (18a) suggests the model performs well under low-clutter conditions, detecting even a small, isolated parasite egg with precision. In the Toxocara example (18b), the detection occurs in a complex and noisy background with numerous debris and artifacts. The model remains accurate even at the edges of the image, showing good spatial generalization across the entire

frame. The *Cystoisospora* predictions in (18c display high localization accuracy despite the image having been captured with a smartphone camera. Similarly, the model succeeds even in the densely populated image with numerous overlapping Giardia cysts (18d), even near the edges of the lens.

The diversity in lighting, image resolution, parasite positioning and image capturing methods, demonstrates the model's ability to generalize well across real-world variations. Detecting in both sparse and cluttered environments shows promise for practical deployment in veterinary diagnostics.

Figure 19 presents an example of a false positive detection among others, it being the left-most one of the objects. The object resembles the approximate round/oval shape and edge contrast of *Toxocara*, possibly confusing the model. Notably, the model assigned a lower confidence score to this detection, suggesting uncertainty.

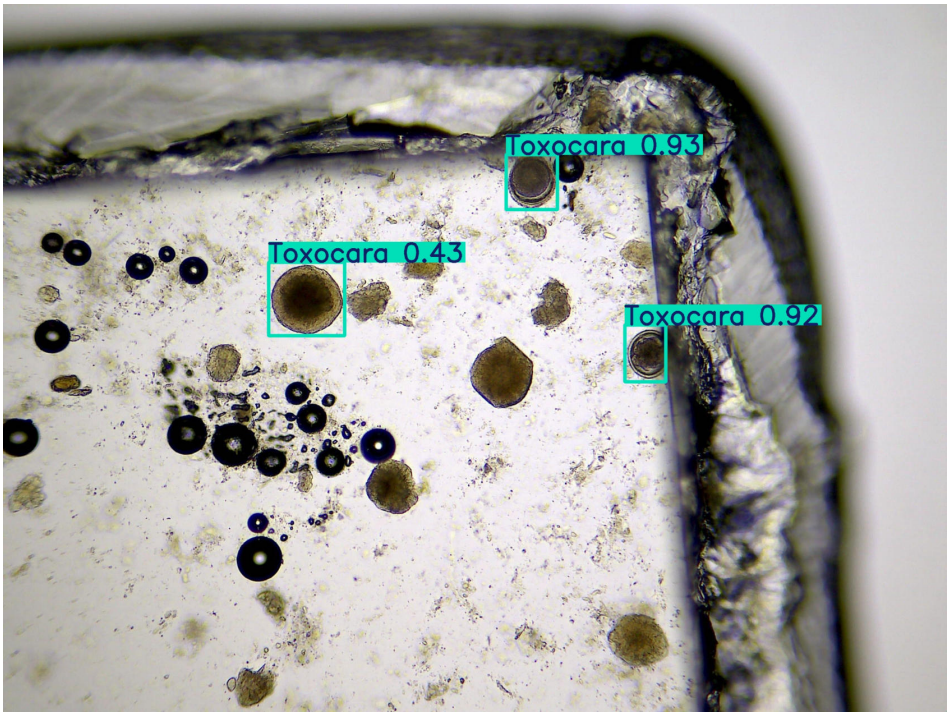


Figure 19: Example of a false positive detection. The object is morphologically similar to the Toxocara, possibly confusing the model.

5.4 Explainability

To evaluate the interpretability of the trained detector in a medical microscopy context, EigenCAM [126] was employed to visualize class-specific saliency overlaid on original input samples. EigenCam is a class activation mapping technique used to visualize and interpret decisions made by CNNs, by highlighting the most influential regions of an input image. It is based on eigenvalue decomposition of feature maps to generate saliency maps without requiring gradient computations or architectural modifications.The figures below present different representative cases across the four parasite species, illustrating both the raw detection output (left) and the corresponding EigenCAM heatmap (right).

Toxocara Figure 20 presents a single Toxocara egg in the center, with a high-confidence detection. The corresponding EigenCAM reveals a sharply focused hotspot directly around the predicted object, with peripheral areas exhibiting low saliency. The EigenCAM localization is highly concentrated and nearly circular, centered precisely over the object. This indicates strong class-discriminative feature activation and minimal contextual dependency. Peripheral suppression shows that the model's feature extraction layers are not distracted by unrelated textures. This sample exemplifies ideal interpretability, where model saliency aligns perfectly with ground truth, affirming its reliability for low-density sample analysis.

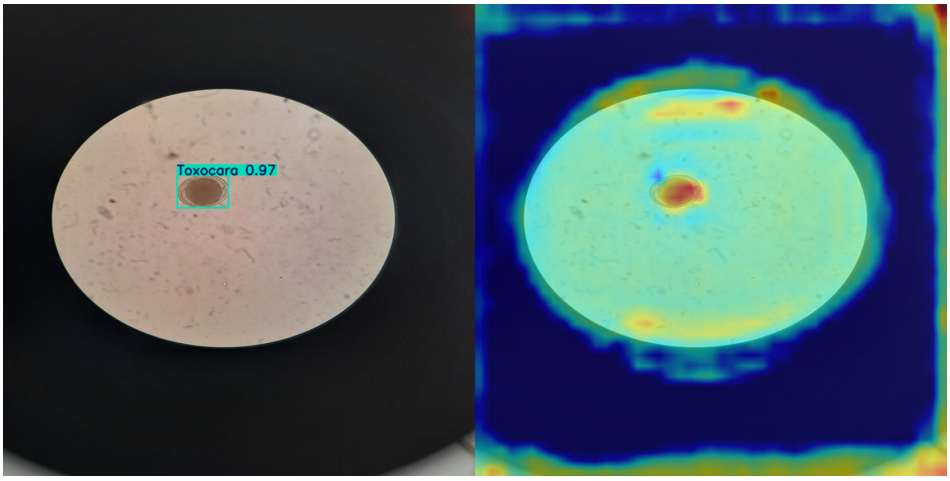


Figure 20: Results using EigenCAM on a single Toxocara egg sample. Left: YOLOv11 detections; Right: Corresponding EigenCAM heatmaps. A sharp hotspot is centered precisely over the object.

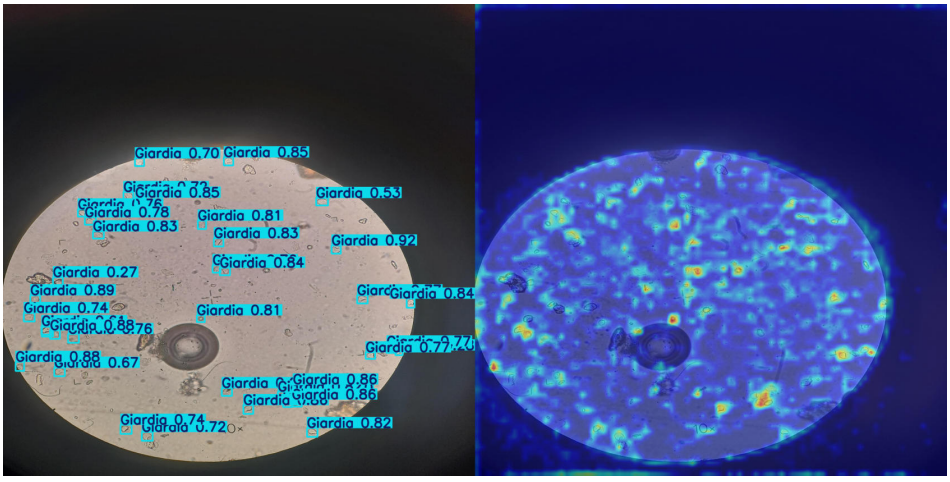


Figure 21: Results using EigenCAM on Giardia cyst samples. Left: YOLOv11 detections; Right: Corresponding EigenCAM heatmaps. The activations are distributed around the relative location of the objects. Background regions remain low-intensity.

Giardia As shown in Figure 21, the model identifies a dense cluster of Giardia cysts with relatively high confidence. The EigenCAM overlay reveals distributed activations with multiple focal hotspots coinciding with the bounding boxes. Despite the presence of a large number of ground truth objects, the attention map does not overly saturate the slide. This suggests the model does not suffer from over-attending or background noise bias. The bright regions (reds and yellows) on the CAM moderately coincide with the locations of predicted bounding boxes, indicating that the model's activations are aligned with the spatial distribution of Giardia instances. Activations in non-target regions remain low-intensity, implying effective spatial discrimination. All in all, this image supports the model's robustness in detecting multiple small, similarly shaped objects and justifies its high confidence scores via interpretable attention.

Cystoisospora Figure 22 involves three clearly bounded instances of Cystoisospora, each confidently detected. The EigenCAM overlay reflects highly localized and intense saliency on these three regions, with a notable absence of significant activation elsewhere. CAM peaks precisely overlap with the prediction boxes, suggesting that feature learning has succeeded in isolating key morphological signatures of Cystoisospora. Unlike the Giardia sample, the heatmap is sparse, with minimal noise. This supports the model's capability to isolate and differentiate sparse objects in low-clutter fields. This visualization strengthens the case for the model's localization ability in detecting non-clustered, distinct parasitic features with minimal background confusion.

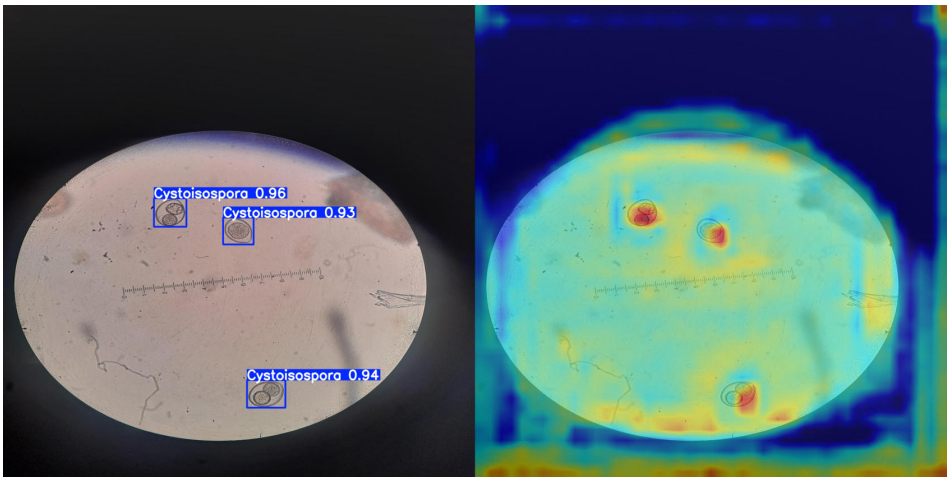


Figure 22: Results using EigenCAM on Cystoisospora samples. Left: YOLOv11 detections; Right: Corresponding EigenCAM heatmaps. The objects are confidently detected. The noise remains low.

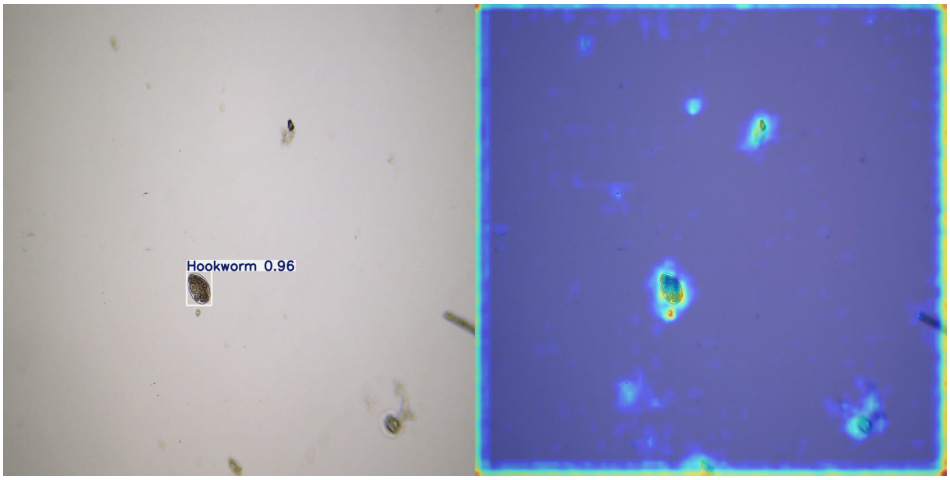


Figure 23: Results using EigenCAM on a Hookworm sample. Left: YOLOv11 detections; Right: Corresponding EigenCAM heatmaps. The heatmap activation is aligned with the boundaries of the object. Some background noise is evident due to possible morphological similarities which however did not lead to a false detection.

Hookworm Finally, in Figure 23, a visualization of a model prediction corresponding to a sample containing a Hookworm ova. Once again, The EigenCAM output demonstrates high spatial precision: the strongest activations are tightly localized around the annotated egg, indicating that the model's confidence is driven by salient morphological features within the egg boundary. This alignment validates the semantic focus of the network and suggests that the learned representations encapsulate relevant parasitological cues. In addition to the primary region, a secondary activation is observed in the upper-middle quadrant. While this region lacks a detection label, its activation implies latent visual similarity to the target class, possibly due to partial structural resemblance.

Across all cases, EigenCAM visualizations corroborate the spatial alignment between predicted detections and model attention. Such correspondence validates the internal consistency of the detector and strengthens its applicability in diagnostic or high-risk environments. Moreover, the absence of significant saliency in false-positive regions suggests a low activation bias in the learned representations.

6 Conclusion and Future Work

Accurate detection of intestinal parasites in companion animals, which can enable timely treatment and reduce the risk of zoonotic transmission, remains a persistent challenge. Early and reliable diagnosis is critical not only for improving animal welfare but also for safeguarding public health and mitigating the economic costs of misdiagnosis or delayed intervention. Traditional diagnostic techniques such as fecal flotation, microscopy, and manual species identification by trained parasitologists, while effective, are time-consuming, labor-intensive, and susceptible to inter-observer variability. In this thesis, the application of computer vision was explored as a solution for automating and improving the accuracy of parasite detection in fecal microscopy images of dogs and cats. This section summarizes the key findings of the research, discusses its methodological and practical limitations, and outlines opportunities for further study and real-world implementation.

6.1 Conclusions

This thesis has presented a comprehensive investigation into the application of modern computer vision techniques for the automated detection of intestinal parasites in microscopy images of companion animals, specifically dogs and cats. The motivation behind this work stems from the persistent diagnostic challenges in veterinary parasitology, where accurate and timely identification of parasitic infections remains crucial for animal welfare and public health. In light of these challenges, this study explored how object detection architectures, particularly those capable of real-time inference, can be adapted and optimized for microscopic diagnostic workflows.

Building upon foundational concepts in deep learning and object detection, the study proposed a pipeline centered around a YOLO-based architecture, leveraging its recent advances. Given the limited prior research on this specific problem, significant effort was allocated to data collection and object detection techniques. The pipeline was trained and validated on a curated dataset consisting of high-resolution microscopy images carefully annotated with the supervision and guidance of a domain expert, covering multiple parasite species.

Quantitative evaluations on validation and test sets demonstrated that the proposed system achieved high detection metrics overall, with strong localization accuracy, and minimal false positive rates across classes. Furthermore, attention-based visualizations confirmed that the network learned salient morphological cues associated with parasitic forms, thus reinforcing the interpretability of predictions—an essential criterion for any model intended for medical deployment.

A key distinction of this work lies in its emphasis on real-time performance and deployability. Unlike traditional CNN-based classifiers or post-processing-heavy segmentation pipelines, the object detection approach used here balances speed and accuracy, making it well-suited for integration into point-of-care diagnostic devices. This capability is particularly important in low-resource or field settings, where expert parasitologists may not always be available.

6.2 Limitations

While the work presented in this thesis has demonstrated encouraging results in the development of automated computer vision techniques for the detection and classification of intestinal parasites in companion animals, several limitations must be acknowledged. These limitations reflect both the practical constraints of the current study and broader challenges in applying AI-driven diagnostic tools in veterinary parasitology. Recognizing these constraints is essential to interpreting the results accurately and guiding future research toward more robust and generalizable systems.

One of the principal limitations encountered in this study was the limited size and relative class imbalance of the dataset. Although the dataset included multiple parasite species across a range of samples, the frequency distribution of classes was uneven, with specifically the *Hookworm* species being significantly underrepresented. Additionally, the overall dataset size restricted the complexity of models that could be trained effectively without overfitting, especially given the high intra-class variability in parasite morphology and imaging artifacts.

Another key limitation was the lack of comprehensive representation of staining protocols and sample preparation methods. Variations in staining intensity, color contrast, and debris contamination across microscopy slides can dramatically affect the appearance of parasitic structures. The dataset may not sufficiently capture the full spectrum of such variations encountered in routine diagnostic laboratories, limiting the model's generalization capacity to unseen samples. Similarly, imaging equipment heterogeneity—such as differences in magnification levels, camera sensors, and lighting conditions—can further contribute to domain shifts not accounted for during model training.

Finally, computational resource constraints presented a significant bottleneck in this work. Due to limitations in available GPU memory and processing time, the scale and diversity of training experiments, including hyperparameter tuning, had to be restricted. As a result, the exploration of alternative model architectures, more complex augmentation pipelines, or ensemble approaches was not feasible within the scope of this thesis.

6.3 Future Work

The present study has demonstrated the feasibility and promise of using deep learning and computer vision techniques for the automated detection of intestinal parasites in microscopy images. In doing so, it opens numerous avenues for future exploration. Advancing this work from a research prototype to a clinically viable diagnostic tool will require addressing several dimensions of development, including dataset expansion, model innovation, clinical validation, deployment infrastructure, interface design, domain adaptation, and ethical oversight. These directions are discussed in detail below.

Dataset Expansion and Parasite Diversity A primary avenue for future research is the substantial expansion and diversification of the dataset. The current study is limited by the relatively small number of labeled microscopy images and the uneven representation of different parasite species.

Increasing the volume of annotated data will allow the model to generalize more robustly and improve performance across both common and rare classes.

Moreover, future datasets should aim to include a broader spectrum of intestinal parasites, with additional ones that infect dogs and cats, but also expanding to other animals. Collaborative data sharing between veterinary institutions, diagnostic laboratories, and research groups could facilitate this expansion and ensure representativeness across geographies and sample preparation methods.

Clinical Integration and Real-World Validation A crucial next step involves evaluating the system in real clinical or laboratory settings, where variables such as sample quality, technician variability, and hardware differences become significant. Conducting prospective clinical trials in veterinary diagnostic laboratories will help assess the model's practical reliability and effectiveness under realistic conditions.

This also includes measuring diagnostic time saved, reduction in observer variability, and overall acceptance by veterinary professionals. User studies could be conducted to assess the interpretability of the results and gauge confidence in the AI-assisted workflow.

Hardware Deployment and User Interface Development To enable practical usage in point-of-care settings or low-resource environments, the model should be deployed on compact, cost-effective hardware platforms. This will involve optimizing the model for inference efficiency, power consumption, and memory footprint.

In parallel, a graphical user interface should be developed to allow seamless interaction with the system. The interface should include intuitive features such as real-time detection, manual image review, heatmap overlays for explainability and storage/export functions for reporting.

Regulatory and Ethical Considerations Finally, as the model approaches potential real-world use, regulatory and ethical concerns must be proactively addressed. This includes ensuring compliance with local and international standards for AI in medical diagnostics. Moreover, the ethical implications of diagnostic automation, including the role of the veterinarian in the loop, accountability for errors, and equitable access to the technology, must be carefully studied and incorporated into the system's design and dissemination strategy.

By addressing these avenues, future research can transform the current proof-of-concept into a robust, reliable, and ethically sound diagnostic tool. Such advancements will not only support veterinarians in detecting intestinal parasites more accurately and efficiently but also contribute to broader public health efforts in the surveillance and control of zoonotic diseases.

References

[1] J. K. Reaser, E. E. Clark Jr, and N. M. Meyers. "All Creatures Great and Minute: A Public Policy Primer for Companion Animal Zoonoses". In: *Zoonoses and Public Health* 55.8-10 (2008), pp. 385–401. DOI: https://doi.org/10.1111/j.1863-2378.2008.01123.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1863-2378.2008.01123.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1863-2378.2008.01123.x.

[2] Michael Paul, Lonnie King, and Ellen P. Carlin. "Zoonoses of people and their pets: a US perspective on significant pet-associated parasitic diseases". In: *Trends in Parasitology* (2010). DOI: 10.1016/j.pt.2010.01.008. URL: https://doi.org/10.1016/j.pt.2010.01.008.

[3] Peter Deplazes et al. "Role of pet dogs and cats in the transmission of helminthic zoonoses in Europe, with a focus on echinococcosis and toxocarosis". In: *Veterinary Parasitology* 182.1 (2011). Special issue: Zoonoses in a Changing World, pp. 41–53. ISSN: 0304-4017. DOI: https://doi.org/10.1016/j.vetpar.2011.07.014. URL: https://www.sciencedirect.com/science/article/pii/S0304401711004833.

[4] Despoina Kostopoulou et al. "Abundance, zoonotic potential and risk factors of intestinal parasitism amongst dog and cat populations: The scenario of Crete, Greece". In: *Parasites & Vectors* (2017). URL: https://doi.org/10.1186/s13071-017-1989-8.

[5] Jia Chen et al. "Canine and feline parasitic zoonoses in China". In: *Parasites & Vectors* (2012). URL: https://doi.org/10.1186/1756-3305-5-152.

[6] WHO Expert Committee on Prevention and Control of Intestinal Parasitic Infections. *Prevention and Control of Intestinal Parasitic Infections: Report of a WHO Expert Committee*. Technical report series. World Health Organization, 1987. ISBN: 9789241207492. URL: https://books.google.gr/books?id=tneBHjgld5oC.

[7] Giulia Simonato et al. "Contamination of Italian parks with canine helminth eggs and health risk perception of the public". In: *Preventive Veterinary Medicine* 172 (2019), p. 104788. ISSN: 0167-5877. DOI: https://doi.org/10.1016/j.prevetmed.2019.104788. URL: https://www.sciencedirect.com/science/article/pii/S0167587719304489.

[8] Dieter Barutzki and Roland Schaper. "Endoparasites in dogs and cats in Germany 1999-2002". In: *Parasitology research* 90 Suppl 3 (July 2003), S148–50. DOI: 10.1007/s00436-003-0922-6.

[9] Frédéric Beugnet et al. "Parasites of domestic owned cats in Europe: co-infestations and risk factors". In: *Parasites & Vectors* (2014). URL: https://doi.org/10.1186/1756-3305-7-291.

[10] Marwan Osman et al. "Prevalence and genetic diversity of the intestinal parasites Blastocystis sp. and Cryptosporidium spp. in household dogs in France and evaluation of zoonotic transmission risk". In: *Veterinary Parasitology* 214.1 (2015), pp. 167–170. ISSN: 0304-4017. DOI: https://doi.org/10.1016/j.vetpar.2015.09.015. URL: https://www.sciencedirect.com/science/article/pii/S0304401715300248.

[11] Alain Villeneuve et al. “Parasite prevalence in fecal samples from shelter dogs and cats across the Canadian provinces”. In: *Parasites & Vectors* (2015). URL: https://doi.org/10.1186/s13071-015-0870-x.

[12] Kristen R. Hoggard et al. “Prevalence survey of gastrointestinal and respiratory parasites of shelter cats in northeastern Georgia, USA”. In: *Veterinary Parasitology: Regional Studies and Reports* 16 (2019), p. 100270. ISSN: 2405-9390. DOI: https://doi.org/10.1016/j.vprsr.2019.100270. URL: https://www.sciencedirect.com/science/article/pii/S2405939018302272.

[13] Yoko Nagamori et al. “Retrospective survey of endoparasitism identified in feces of client-owned dogs in North America from 2007 through 2018”. In: *Veterinary Parasitology* 282 (2020), p. 109137. ISSN: 0304-4017. DOI: https://doi.org/10.1016/j.vetpar.2020.109137. URL: https://www.sciencedirect.com/science/article/pii/S0304401720301175.

[14] Josef Finsterer and Herbert Auer. “Neurotoxocarosis”. In: *Revista do Instituto de Medicina Tropical de São Paulo* 49.5 (2007), 279–287. ISSN: 0036-4665. DOI: 10.1590/S0036-46652007000500002. URL: https://doi.org/10.1590/S0036-46652007000500002.

[15] Huw Smith et al. “How common is human toxocariasis? Towards standardizing our knowledge”. In: *Trends in Parasitology* (2009). DOI: 10.1016/j.pt.2009.01.006. URL: https://doi.org/10.1016/j.pt.2009.01.006.

[16] G. Rubinsky-Elefant et al. “Human toxocariasis: diagnosis, worldwide seroprevalences and clinical expression of the systemic and ocular forms”. In: *Annals of Tropical Medicine & Parasitology* 104.1 (2010). PMID: 20149289, pp. 3–23. DOI: 10.1179/136485910X12607012373957. eprint: https://doi.org/10.1179/136485910X12607012373957. URL: https://doi.org/10.1179/136485910X12607012373957.

[17] Rebecca J. Traub. “Ancylostoma ceylanicum, a re-emerging but neglected parasitic zoonosis”. In: *International Journal for Parasitology* 43.12 (2013). Zoonoses Special Issue, pp. 1009–1015. ISSN: 0020-7519. DOI: https://doi.org/10.1016/j.ijpara.2013.07.006. URL: https://www.sciencedirect.com/science/article/pii/S0020751913002038.

[18] Dwight D. Bowman et al. “Hookworms of dogs and cats as agents of cutaneous larva migrans”. In: *Trends in Parasitology* (2010). DOI: 10.1016/j.pt.2010.01.005. URL: https://doi.org/10.1016/j.pt.2010.01.005.

[19] Bernard Nkrumah and Samuel Blay Nguah. “Giardia lamblia: a major parasitic cause of childhood diarrhoea in patients attending a district hospital in Ghana”. In: *Parasites & Vectors* (2011). DOI: 10.1186/1756-3305-4-163. URL: https://doi.org/10.1186/1756-3305-4-163.

[20] Jingjing Sun et al. “Assessment of potential zoonotic transmission of Giardia duodenalis from dogs and cats”. In: *One Health* 17 (2023), p. 100651. ISSN: 2352-7714. DOI: https://doi.org/10.1016/j.onehlt.2023.100651. URL: https://www.sciencedirect.com/science/article/pii/S2352771423001714.

[21] IL Mitrea, Mariana IoniŃă, and Violeta Enăchescu. “ISOSPOROSIS IN CATS AND DOGS: ETIO-EPIDEMIOLOGICAL AND CLINICAL FEATURES”. In: *Buletin USAMV-CN* 63 (2006), pp. 343–347. URL: https://ftp.fruit-technology.ro/index.php/veterinary/article/view/2508.

[22] Anthony Andrews. *Coccidiosis of Cats and Dogs*. Web Page. 2022. URL: https://www.msdvetmanual.com/digestive-system/coccidiosis/coccidiosis-of-cats-and-dogs.

[23] JD Broussard. “Optimal fecal assessment”. In: *Clinical Techniques in Small Animal Practice* 18 (Nov. 2003), pp. 218–230. DOI: 10.1053/S1906-2867(03)00076-8.

[24] Noppadon Tangpukdee et al. “Malaria Diagnosis: A Brief Review”. In: *Korean Journal of Parasitology* 47.2 (2009), pp. 93–102. DOI: 10.3347/kjp.2009.47.2.93. eprint: https://www.parahostdis.org/journal/view.php?number=118. URL: https://www.parahostdis.org/journal/view.php?number=118.

[25] Susan E. Little et al. “Prevalence of intestinal parasites in pet dogs in the United States”. In: *Veterinary Parasitology* 166.1 (2009), pp. 144–152. ISSN: 0304-4017. DOI: https://doi.org/10.1016/j.vetpar.2009.07.044. URL: https://www.sciencedirect.com/science/article/pii/S030440170900466X.

[26] Sen Li et al. “Transfer Learning for Toxoplasma gondii Recognition”. In: *mSystems* 5.1 (2020), 10.1128/msystems.00445–19. DOI: 10.1128/msystems.00445-19. eprint: https://journals.asm.org/doi/pdf/10.1128/msystems.00445-19. URL: https://journals.asm.org/doi/abs/10.1128/msystems.00445-19.

[27] Maciej A. Mazurowski et al. *Deep learning in radiology: an overview of the concepts and a survey of the state of the art*. 2018. arXiv: 1802.08717 [cs.CV]. URL: https://arxiv.org/abs/1802.08717.

[28] Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. “Deep neural network models for computational histopathology: A survey”. In: *Medical Image Analysis* 67 (2021), p. 101813. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2020.101813. URL: https://www.sciencedirect.com/science/article/pii/S1361841520301778.

[29] Seung Seog Han et al. “Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network”. In: *PLOS ONE* 13.1 (Jan. 2018), pp. 1–14. DOI: 10.1371/journal.pone.0191493. URL: https://doi.org/10.1371/journal.pone.0191493.

[30] Jeffrey Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". In: *Nature Medicine* 24 (Sept. 2018). DOI: 10.1038/s41591-018-0107-6.

[31] Parampal Grewal et al. "Deep learning in ophthalmology: a review". In: *Canadian Journal of Ophthalmology* 53 (May 2018). DOI: 10.1016/j.jcjo.2018.04.019.

[32] John A. Quinn et al. *Deep Convolutional Neural Networks for Microscopy-Based Point of Care Diagnostics*. 2016. arXiv: 1608.02989 [cs.CV]. URL: https://arxiv.org/abs/1608.02989.

[33] S. Kevin Zhou et al. "A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises". In: *Proceedings of the IEEE* 109.5 (May 2021), 820–838. ISSN: 1558-2256. DOI: 10.1109/jproc.2021.3054390. URL: http://dx.doi.org/10.1109/JPROC.2021.3054390.

[34] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2017.07.005. URL: https://www.sciencedirect.com/science/article/pii/S1361841517301135.

[35] Fuyong Xing et al. "Deep Learning in Microscopy Image Analysis: A Survey". In: *IEEE Transactions on Neural Networks and Learning Systems* 29.10 (2018), pp. 4550–4568. DOI: 10.1109/TNNLS.2017.2766168.

[36] Andre Esteva et al. "A guide to deep learning in healthcare". In: *Nature Medicine* 25 (Jan. 2019). DOI: 10.1038/s41591-018-0316-z.

[37] Massimo Salvi et al. "The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis". In: *Computers in Biology and Medicine* 128 (2021), p. 104129. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2020.104129. URL: https://www.sciencedirect.com/science/article/pii/S0010482520304601.

[38] Subhash Chandra Parija and Abhijit Poddar. "Artificial intelligence in parasitic disease control: A paradigm shift in health care". In: *Tropical Parasitology* (2024). DOI: 10.4103/tp.tp_66_23. URL: https://journals.lww.com/tpar/fulltext/2024/14010/artificial_intelligence_in_parasitic_disease.2.aspx.

[39] J Redmon. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[40] Haoyu Liu. "Systematic study of lightweight for object detection". In: *Applied and Computational Engineering* 73 (July 2024), pp. 8–15. DOI: 10.54254/2755-2721/73/20240354.

[41] Dingjun Yu et al. "Mixed Pooling for Convolutional Neural Networks". In: Oct. 2014, pp. 364–375. ISBN: 978-3-319-11739-3. DOI: 10.1007/978-3-319-11740-9_34.

[42] Priyanto Hidayatullah et al. *YOLOv8 to YOLO11: A Comprehensive Architecture In-depth Comparative Review*. 2025. arXiv: 2501.13400 [cs.CV]. URL: https://arxiv.org/abs/2501.13400.

[43] Paul Viola and Michael Jones. "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Vol. 1. Ieee. 2001, pp. I–I.

[44] Pedro F Felzenszwalb et al. "Object detection with discriminatively trained part-based models". In: *IEEE transactions on pattern analysis and machine intelligence* 32.9 (2009), pp. 1627–1645.

[45] Radomir S Stankovic and Bogdan J Falkowski. "The Haar wavelet transform: its status and achievements". In: *Computers & Electrical Engineering* 29.1 (2003), pp. 25–44.

[46] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. Ieee. 2005, pp. 886–893.

[47] Zhengxia Zou et al. "Object detection in 20 years: A survey". In: *Proceedings of the IEEE* 111.3 (2023), pp. 257–276.

[48] Ross Girshick et al. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[49] R Girshick. "Fast r-cnn". In: *arXiv preprint arXiv:1504.08083* (2015).

[50] Shaoqing Ren et al. "Faster R-CNN: Towards real-time object detection with region proposal networks". In: *IEEE transactions on pattern analysis and machine intelligence* 39.6 (2016), pp. 1137–1149.

[51] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[52] Wei Liu et al. "Ssd: Single shot multibox detector". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 21–37.

[53] T Lin. "Focal Loss for Dense Object Detection". In: *arXiv preprint arXiv:1708.02002* (2017).

[54] Nicolas Carion et al. "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[55] Xizhou Zhu et al. "Deformable detr: Deformable transformers for end-to-end object detection". In: *arXiv preprint arXiv:2010.04159* (2020).

[56] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 7464–7475.

[57] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. *Ultralytics YOLOv8*. Version 8.0.0. 2023. URL: https://github.com/ultralytics/ultralytics.

[58] Hao Zhang et al. "Dino: Detr with improved denoising anchor boxes for end-to-end object detection". In: *arXiv preprint arXiv:2203.03605* (2022).

[59] Juan Terven, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. "A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas". In: *Machine Learning and Knowledge Extraction* 5.4 (2023), pp. 1680–1716.

[60] Alexander Kirillov et al. "Segment anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[61] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Commun. ACM* 60.6 (May 2017), 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386. URL: https://doi.org/10.1145/3065386.

[62] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV]. URL: https://arxiv.org/abs/1409.1556.

[63] Christian Szegedy et al. *Going Deeper with Convolutions*. 2014. arXiv: 1409.4842 [cs.CV]. URL: https://arxiv.org/abs/1409.4842.

[64] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[65] Mark Sandler et al. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. 2019. arXiv: 1801.04381 [cs.CV]. URL: https://arxiv.org/abs/1801.04381.

[66] Mingxing Tan and Quoc V. Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. 2020. arXiv: 1905.11946 [cs.LG]. URL: https://arxiv.org/abs/1905.11946.

[67] Mingxing Tan and Quoc V. Le. *EfficientNetV2: Smaller Models and Faster Training*. 2021. arXiv: 2104.00298 [cs.CV]. URL: https://arxiv.org/abs/2104.00298.

[68] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

[69] Hugo Touvron et al. "Training data-efficient image transformers & distillation through attention". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 10347–10357. URL: https://proceedings.mlr.press/v139/touvron21a.html.

[70] Ze Liu et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 9992–10002. DOI: 10.1109/ICCV48922.2021.00986.

[71] Zhuang Liu et al. "A ConvNet for the 2020s". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 11966–11976. DOI: 10.1109/CVPR52688.2022.01167.

[72] Zihang Dai et al. *CoAtNet: Marrying Convolution and Attention for All Data Sizes*. 2021. arXiv: 2106.04803 [cs.CV]. URL: https://arxiv.org/abs/2106.04803.

[73] Jing Liu, Huiyang Chen, and Weimin Zhou. "Improved MobileViT: A More Efficient Lightweight Convolution and Vision Transformer Hybrid Model". In: *Journal of Physics: Conference Series* 2562.1 (2023), p. 012012. DOI: 10.1088/1742-6596/2562/1/012012. URL: https://dx.doi.org/10.1088/1742-6596/2562/1/012012.

[74] Kaiming He et al. *Masked Autoencoders Are Scalable Vision Learners*. 2021. arXiv: 2111.06377 [cs.CV]. URL: https://arxiv.org/abs/2111.06377.

[75] Yuxin Fang et al. *EVA: Exploring the Limits of Masked Visual Representation Learning at Scale*. 2022. arXiv: 2211.07636 [cs.CV]. URL: https://arxiv.org/abs/2211.07636.

[76] Yoon Seok Yang et al. "Automatic identification of human helminth eggs on microscopic fecal specimens using digital image processing and an artificial neural network". In: *IEEE Transactions on Biomedical Engineering* 48.6 (2001), pp. 718–730. DOI: 10.1109/10.923789.

[77] Kenneth W. Widmer, Kevin H. Oshima, and Suresh D. Pillai. "Identification of ¡i¿Cryptosporidium¡i¿ parvum Oocysts by an Artificial Neural Network Approach". In: *Applied and Environmental Microbiology* 68.3 (2002), pp. 1115–1121. DOI: 10.1128/AEM.68.3.1115-1121.2002. eprint: https://journals.asm.org/doi/pdf/10.1128/aem.68.3.1115-1121.2002. URL: https://journals.asm.org/doi/abs/10.1128/aem.68.3.1115-1121.2002.

[78] Kenneth W. Widmer, Deepak Srikumar, and Suresh D. Pillai. "Use of Artificial Neural Networks To Accurately Identify ¡i¿Cryptosporidium¡/i¿ Oocyst and ¡i¿Giardia¡/i¿ Cyst Images". In: *Applied and Environmental Microbiology* 71.1 (2005), pp. 80–84. DOI: 10.1128/AEM.71.1.80-84.2005. eprint: https://journals.asm.org/doi/pdf/10.1128/aem.71.1.80-84.2005. URL: https://journals.asm.org/doi/abs/10.1128/aem.71.1.80-84.2005.

[79] Esin Dogantekin et al. "A robust technique based on invariant moments – ANFIS for recognition of human parasite eggs in microscopic images". In: *Expert Systems with Applications* 35.3 (2008), pp. 728–738. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2007.07.020. URL: https://www.sciencedirect.com/science/article/pii/S0957417407002862.

[80] Derya Avci and Asaf Varol. "An expert diagnosis system for classification of human parasite eggs based on multi-class SVM". In: *Expert Systems with Applications* 36.1 (2009), pp. 43–48. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2007.09.012. URL: https://www.sciencedirect.com/science/article/pii/S0957417407004538.

[81] Celso T. N. Suzuki et al. “Automatic Segmentation and Classification of Human Intestinal Parasites From Microscopy Images”. In: *IEEE Transactions on Biomedical Engineering* 60.3 (2013), pp. 803–812. DOI: 10.1109/TBME.2012.2187204.

[82] Alan Peixinho. “Diagnosis of Human Intestinal Parasites by Deep Learning”. In: Oct. 2015. DOI: 10.1201/b19241-19.

[83] Beaudelaire Saha Tchinda et al. “Towards an automated medical diagnosis system for intestinal parasitosis”. In: *Informatics in Medicine Unlocked* 13 (2018), pp. 101–111. ISSN: 2352-9148. DOI: https://doi.org/10.1016/j.imu.2018.09.004. URL: https://www.sciencedirect.com/science/article/pii/S2352914818301588.

[84] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV]. URL: https://arxiv.org/abs/1505.04597.

[85] Yaning Li et al. “A low-cost, automated parasite diagnostic system via a portable, robotic microscope and deep learning”. In: *Journal of Biophotonics* 12.9 (2019), e201800410. DOI: https://doi.org/10.1002/jbio.201800410. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jbio.201800410. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jbio.201800410.

[86] Shaoqing Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: 1506.01497 [cs.CV]. URL: https://arxiv.org/abs/1506.01497.

[87] Ngo Quoc Viet, Dang Thi ThanhTuyen, and Trinh Huy Hoang. “Parasite worm egg automatic detection in microscopy stool image based on Faster R-CNN”. In: *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*. ICMLSC ’19. Da Lat, Viet Nam: Association for Computing Machinery, 2019, 197–202. ISBN: 9781450366120. DOI: 10.1145/3310986.3311014. URL: https://doi.org/10.1145/3310986.3311014.

[88] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. arXiv: 1502.03167 [cs.LG]. URL: https://arxiv.org/abs/1502.03167.

[89] Blaine A. Mathison et al. “Detection of Intestinal Protozoa in Trichrome-Stained Stool Specimens by Use of a Deep Convolutional Neural Network”. In: *Journal of Clinical Microbiology* 58.6 (2020), 10.1128/jcm.02053–19. DOI: 10.1128/jcm.02053-19. eprint: https://journals.asm.org/doi/pdf/10.1128/jcm.02053-19. URL: https://journals.asm.org/doi/abs/10.1128/jcm.02053-19.

[90] Joseph Redmon and Ali Farhadi. *YOLOv3: An Incremental Improvement*. 2018. arXiv: 1804.02767 [cs.CV]. URL: https://arxiv.org/abs/1804.02767.

[91] Kristofer E. delas Peñas et al. "Automated Detection of Helminth Eggs in Stool Samples Using Convolutional Neural Networks". In: *2020 IEEE REGION 10 CONFERENCE (TENCON)*. 2020, pp. 750–755. DOI: 10.1109/TENCON50793.2020.9293746.

[92] Narut Butploy, Wanida Kanarkard, and Pewpan Maleewong Intapan. "Deep Learning Approach for Ascaris lumbricoides Parasite Egg Classification". In: *Journal of Parasitology Research* 2021.1 (2021), p. 6648038. DOI: https://doi.org/10.1155/2021/6648038. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/6648038. URL: https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/6648038.

[93] Rose Nakasi, Ezra Rwakazooba Aliija, and Joyce Nakatumba. "A Poster on Intestinal Parasite Detection in Stool Sample Using AlexNet and GoogleNet Architectures". In: *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*. COMPASS '21. Virtual Event, Australia: Association for Computing Machinery, 2021, 389–395. ISBN: 9781450384537. DOI: 10.1145/3460112.3472309. URL: https://doi.org/10.1145/3460112.3472309.

[94] Thanaphon Suwannaphong et al. *Parasitic Egg Detection and Classification in Low-cost Microscopic Images using Transfer Learning*. 2021. arXiv: 2107.00968 [cs.CV]. URL: https://arxiv.org/abs/2107.00968.

[95] Tsung-Yi Lin et al. "Focal Loss for Dense Object Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2 (2020), pp. 318–327. DOI: 10.1109/TPAMI.2018.2858826.

[96] Yibo Huo et al. "Recognition of parasite eggs in microscopic medical images based on YOLOv5". English. In: *Proceedings of 2021 5th Asian Conference on Artificial Intelligence Technology, ACAIT 2021*. Proceedings of 2021 5th Asian Conference on Artificial Intelligence Technology, ACAIT 2021. Publisher Copyright: © 2021 IEEE.; 5th Asian Conference on Artificial Intelligence Technology, ACAIT 2021 ; Conference date: 29-10-2021 Through 31-10-2021. United States: Institute of Electrical and Electronics Engineers Inc., 2021, pp. 123–127. DOI: 10.1109/ACAIT53529.2021.9731120.

[97] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. *YOLOv4: Optimal Speed and Accuracy of Object Detection*. 2020. arXiv: 2004.10934 [cs.CV]. URL: https://arxiv.org/abs/2004.10934.

[98] Chien-Yao Wang et al. *CSPNet: A New Backbone that can Enhance Learning Capability of CNN*. 2019. arXiv: 1911.11929 [cs.CV]. URL: https://arxiv.org/abs/1911.11929.

[99] Naing KM et al. "Automatic recognition of parasitic products in stool examination using object detection approach." In: *PeerJ Computer Science* (2022). URL: https://doi.org/10.7717/peerj-cs.1065.

[100] Perla Mayo et al. *Detection of Parasitic Eggs from Microscopy Images and the emergence of a new dataset*. 2022. arXiv: 2203.02940 [cs.CV]. URL: https://arxiv.org/abs/2203.02940.

[101] Mingxing Tan, Ruoming Pang, and Quoc V. Le. *EfficientDet: Scalable and Efficient Object Detection*. 2020. arXiv: 1911.09070 [cs.CV]. URL: https://arxiv.org/abs/1911.09070.

[102] Duangdao Palasuwan; Korranat Naruenatthanaset; Thananop Kobchaisawat; Thanarat H Chalidabhongse; Nuntiporn Nunthanasup; Kanyarat Boonpeng; Nantheera Anantrasirichai. *Parasitic Egg Detection and Classification in Microscopic Images*. 2022. DOI: 10.21227/vyh8-4h71. URL: https://dx.doi.org/10.21227/vyh8-4h71.

[103] Nouar Aldahoul et al. "Localization and Classification of Parasitic Eggs in Microscpic Images Using An Efficientdet Detector". In: Oct. 2022. DOI: 10.1109/ICIP46576.2022.9897844.

[104] Jingdong Wang et al. "Deep High-Resolution Representation Learning for Visual Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.10 (2021), pp. 3349–3364. DOI: 10.1109/TPAMI.2020.2983686.

[105] Saining Xie et al. "Aggregated Residual Transformations for Deep Neural Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5987–5995. DOI: 10.1109/CVPR.2017.634.

[106] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. "Weighted boxes fusion: Ensembling boxes from different object detection models". In: *Image and Vision Computing* 107 (Mar. 2021), p. 104117. ISSN: 0262-8856. DOI: 10.1016/j.imavis.2021.104117. URL: http://dx.doi.org/10.1016/j.imavis.2021.104117.

[107] Zaw Htet Aung, Kittinan Srithaworn, and Titipat Achakulvisut. "Multitask learning via pseudo-label generation and ensemble prediction for parasitic egg cell detection: IEEE ICIP Challenge 2022". In: *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, pp. 4273–4277. DOI: 10.1109/ICIP46576.2022.9897464.

[108] Yuqi Wang et al. "A Robust Ensemble Model For Parasitic Egg Detection And Classification". In: *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, pp. 4258–4262. DOI: 10.1109/ICIP46576.2022.9897192.

[109] Rejin Varghese and Sambath M. "YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness". In: *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*. 2024, pp. 1–6. DOI: 10.1109/ADICS58448.2024.10533619.

[110] Yuxin Wu et al. *Detectron2*. https://github.com/facebookresearch/detectron2. 2019.

[111] Christian Szegedy et al. *Rethinking the Inception Architecture for Computer Vision*. 2015. arXiv: 1512.00567 [cs.CV]. URL: https://arxiv.org/abs/1512.00567.

[112] Yuxin Fang et al. *You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection*. 2021. arXiv: 2106.00666 [cs.CV]. URL: https://arxiv.org/abs/2106.00666.

[113] Sakthi Jaya Sundar Rajasekar et al. "Parasite.ai – An Automated Parasitic Egg Detection Model from Microscopic Images of Fecal Smears using Deep Learning Techniques". In: *2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)* (2023), pp. 1–9. URL: https://api.semanticscholar.org/CorpusID:260513855.

[114] Nouar AlDahoul et al. "Parasitic egg recognition using convolution and attention network". In: *Scientific Reports* (2023). URL: https://doi.org/10.1038/s41598-023-43068-z.

[115] Satish Kumar et al. "An Efficient and Effective Framework for Intestinal Parasite Egg Detection Using YOLOv5". In: *Diagnostics* 13.18 (2023). ISSN: 2075-4418. DOI: 10.3390/diagnostics13182978. URL: https://www.mdpi.com/2075-4418/13/18/2978.

[116] François Chollet. *Xception: Deep Learning with Depthwise Separable Convolutions*. 2017. arXiv: 1610.02357 [cs.CV]. URL: https://arxiv.org/abs/1610.02357.

[117] Andrew G. Howard et al. *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. 2017. arXiv: 1704.04861 [cs.CV]. URL: https://arxiv.org/abs/1704.04861.

[118] Natthanai Chaibutr et al. "Development of a Machine Learning Model for the Classification of Enterobius vermicularis Egg". In: *Journal of Imaging* 10.9 (2024). ISSN: 2313-433X. DOI: 10.3390/jimaging10090212. URL: https://www.mdpi.com/2313-433X/10/9/212.

[119] Sandra Valéria Inácio et al. "Automated Diagnosis of Canine Gastrointestinal Parasites Using Image Analysis". In: *Pathogens* 9.2 (2020). ISSN: 2076-0817. DOI: 10.3390/pathogens9020139. URL: https://www.mdpi.com/2076-0817/9/2/139.

[120] Khaye C. Fajardo, Jerald SD. Gonzales, and Carlos C. Hortinela IV. "Detection and Identification of Intestinal Parasites on Dogs Using AlexNet CNN Architecture". In: *2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*. 2022, pp. 1–6. DOI: 10.1109/IICAIET55139.2022.9936777.

[121] Xuebin Qin et al. "U2-Net: Going deeper with nested U-structure for salient object detection". In: *Pattern Recognition* 106 (2020), p. 107404. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2020.107404. URL: https://www.sciencedirect.com/science/article/pii/S0031320320302077.

[122] Gao Huang et al. "Densely Connected Convolutional Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

[123] L.M. Joao et al. "Toward automating the diagnosis of gastrointestinal parasites in cats and dogs". In: *Computers in Biology and Medicine* 163 (2023), p. 107203. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2023.107203. URL: https://www.sciencedirect.com/science/article/pii/S0010482523006686.

[124] Yoko Nagamori et al. “Multicenter evaluation of the Vetscan Imagyst system using Ocus 40 and EasyScan One scanners to detect gastrointestinal parasites in feces of dogs and cats”. In: *Journal of Veterinary Diagnostic Investigation* 36.1 (2024). PMID: 38014739, pp. 32–40. DOI: 10.1177/10406387231216185. eprint: https://doi.org/10.1177/10406387231216185. URL: https://doi.org/10.1177/10406387231216185.

[125] Tsung-Yi Lin et al. *Microsoft COCO: Common Objects in Context*. 2015. arXiv: 1405.0312 [cs.CV]. URL: https://arxiv.org/abs/1405.0312.

[126] Mohammed Bany Muhammad and Mohammed Yeasin. “Eigen-CAM: Class Activation Map using Principal Components”. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2020, 1–7. DOI: 10.1109/ijcnn48605.2020.9206626. URL: http://dx.doi.org/10.1109/IJCNN48605.2020.9206626.