Wheat Head Detection Using Detection Transformer (DETR)

Hamna Mansoor
Dhanani School of Science & Engineering
Habib University
Karachi, Pakistan

Shehryar Amin
Dhanani School of Science & Engineering
Habib University
Karachi, Pakistan

Abstract

Wheat, a cornerstone of global agriculture, has played a pivotal role in global food security, providing essential nutrients and sustenance to millions worldwide. According to the Food and Agriculture Organization of the United Nations, wheat production reached a staggering 765.76 million tons in 2019 (El-Hendawy et al. 2022), underscoring its indispensable role in meeting the dietary needs of populations across the globe. As an integral component of human civilization, wheat cultivation has undergone profound transformations, transitioning from traditional agricultural practices to the adoption of cutting-edge technologies in modern farming. Amidst this evolution, the field of wheat phenotyping has emerged as a focal point of research, driven by its paramount importance in promoting food security and fostering sustainable agricultural practices.

In this paper, the efficacy of a transformer-based model, specifically the DEtection TRansformer (DETR) architecture proposed by Carion et al. in 2020, for the task of wheat head detection is explored. Firstly, a DETR model is fine-tuned with a ResNet-50 convolutional backbone, renowned for its robust feature extraction capabilities, utilizing the Global Wheat Head Detection dataset (GWHD). Leveraging the ResNet-50 architecture as the backbone enhances the feature extraction process within the DETR framework. Subsequently, a comprehensive performance analysis of the fine-tuned model is conducted, comparing the results against those obtained from existing methodologies, predominantly rooted in YOLO and Faster R-CNN-based approaches. This comparative evaluation seeks to ascertain the effectiveness of the DETR architecture in addressing wheat head detection tasks and elucidate its potential contributions to the field.

Introduction

Wheat phenotyping is an essential practice in agricultural research, crucial for understanding and improving crop performance. It involves the detailed quantification of various traits that influence yield, resilience, and quality in wheat varieties. This includes traits such as:

- Plant height
- Leaf size and shape
- Grain yield
- Seed number, shape, and size
- Biomass production
- Drought tolerance
- Disease resistance
- Vigour
- Spikelet number per spike (SPS)

The accurate quantification of these traits is indispensable for gaining profound insights into plant development and for refining agricultural methodologies (Velu and Singh 2013). This precision ensures that researchers can capture the subtle nuances that differentiate various wheat varieties, enabling more informed decision-making in breeding programs and agronomic strategies. Experimental trials have formed the foundation of phenotypic studies, furnishing invaluable data for elucidating wheat phenotypes and assessing crop performance across diverse environmental conditions. An illustrative case study from South Australia offers a compelling demonstration of meticulous trial design (Batin et al. 2023). In this study, ten varieties of spring wheat underwent rigorous evaluation through meticulously crafted field trials. Employing a randomized split-block design, with distinct plots assigned for fertilized and control conditions, ensured not only the robustness of statistical analyses but also facilitated accurate comparisons of treatment effects. The significance of this approach lies in its ability to capture the nuances of wheat phenotypes under varying experimental conditions. By meticulously recording plant growth throughout the growing season, researchers are afforded the opportunity to delineate the dynamic expression of traits and discern the impact of external factors, such as fertilization regimes, on wheat phenotypes.

Moreover, phenotyping allows breeders to identify and select plants with desirable traits for further breeding. By meticulously observing and analyzing physical characteristics, breeders gain insights into the genetic makeup and performance of different wheat varieties. This understanding allows them to strategically choose parent plants with complementary traits to cross-pollinate, ultimately aiming to develop new wheat varieties that exhibit improved traits, such as higher yields, better resistance to pests and diseases, and enhanced adaptability to diverse environmental conditions (Velu and Singh 2013).

Traditional methods of assessing wheat traits, such as manual counting and observation, are labor-intensive, subjective, and prone to errors. This limitation hampers the breeding process and ultimately impacts wheat yield, posing a challenge to global food security. Studies have highlighted the significant measurement errors—up to 10%—associated with manual counting methods (Tian et al. 2020). As a result, there is a growing need within the agricultural sector to streamline and automate wheat cultivation processes through advanced technological solutions. The significance of this paper lies in its potential to transform wheat phenotyping practices, offering a more accurate, efficient, and automated solution for wheat head detection. This paper focuses on developing an automated system for wheat head detection leveraging DETR, in outdoor field images. The scope encompasses the implementation of the DEtection TRansformer (DETR) architecture proposed by Nicolas Carion. This involves fine-tuning the DETR model with a ResNet-50 backbone for feature extraction and analyzing its performance against existing methods, such as YOLO and Faster R-CNN.

Literature Review

Various methods have been proposed to automate the process of wheat head detection, ranging from traditional image processing techniques to cutting-edge deep learning approaches. One pioneering contribution in this domain was presented by Yangjun Zhu et al. in their paper published in 2016. Their two-step coarse-to-fine wheat-head detection method represents a milestone in automated crop analysis (Zhu et al. 2016). By meticulously delineating wheat heads from the image background through intricate binarization thresholds, Zhu et al. laid the groundwork for subsequent nuanced analysis, offering a robust solution for wheat head detection (Bi et al. 2010).

Similarly, Tang et al. explored wheat head detection based on RGB photos using classical image processing techniques (TANG et al. 2017). Leveraging tools such as the Laplacian frequency filter and median filter, they unveiled a method that surpassed previous benchmarks, achieving a detection accuracy exceeding 90% in their test set. Their innovative approach not only demonstrated the efficacy of traditional image processing methodologies but also underscored the importance of adaptability and innovation in addressing real-world agricultural challenges.

Transitioning from traditional methods to cutting-edge technologies, Uddin et al. presented a groundbreaking study on rice spike assessment, showcasing the transformative potential of Convolutional Neural Networks (CNNs). Their integration of the Feature Pyramid Network (FPN) into the Faster Region-based CNN (Faster R-CNN) architecture yielded remarkable accuracy, nearing 99% (Mia et al. 2021). This fusion of traditional deep learning frameworks with innovative architectural enhancements opened new avenues for precise crop analysis, revolutionizing the way agricultural data is processed and interpreted.

Additionally, Zhu et al. presented a novel approach to wheat head detection using deep learning. The research addresses the critical need for accurate detection methods in the context of challenges faced by wheat supply due to factors such as population growth and climate change. Specifically, existing deep learning models struggle with identifying wheat heads in UAV-captured images characterized by high density and overlapping instances. To overcome these limitations, the study employs high-resolution wheat head images captured by UAVs, providing both temporal and spatial detail. The authors introduce three distinct object detection networks based on Transformer architecture: FR-Transformer, R-Transformer, and Y-Transformer. Among these, the FR-Transformer stands out as a two-stage method, demonstrating superior performance compared to conventional CNN-based approaches, achieving an 88.3% improvement for mAP@50 and 38.5% for mAP@75 (Zhu et al. 2022). By leveraging the Transformer's capability to capture global features, the proposed models enhance detection accuracy while reducing computational complexity. Ultimately, the FR-Transformer method offered a promising solution for the rapid and precise detection of wheat heads in UAV-captured images, addressing the critical need for accurate wheat yield estimation in agricultural practices.

In 2023, the YOLOv7-MA model was employed for wheat head detection and counting, aiming to overcome challenges such as overlapping and small wheat heads against complex backgrounds. The model introduced micro-scale detection layers and a convolutional block attention module to enhance target information for wheat heads while mitigating background noise, resulting in improved detection performance. Trained and evaluated on the Global Wheat Head Dataset 2021, YOLOv7-MA achieved a mean average precision (mAP) of 93.86% with a detection speed of 35.93 frames per second (FPS), surpassing the performance of Faster-RCNN, YOLOv5, YOLOX, and YOLOv7 models (Meng et al. 2023). Notably, YOLOv7-MA demonstrated robustness under challenging conditions such as low illumination, blur, and occlusion, with a stronger correlation between predicted wheat head count and manual counting compared to other models. In field applications using wheat head datasets collected from the field, YOLOv7-MA maintains high performance during maturity and filling stages, suggesting its potential for providing technical support for large-scale wheat yield estimation using unmanned aerial vehicles (UAVs).

In a recent study, researchers proposed the YOLOv8-HD Model for wheat seed detection and counting. The method introduced YOLOv8-HD, a lightweight real-time wheat seed detection model. YOLOv8-HD incorporates improvements such as shared convolutional layers for parameter reduction and a Vision Transformer with a Deformable Attention mechanism for enhanced feature extraction and detection accuracy. Results indicated that YOLOv8-HD outperformed YOLOv8, achieving an average detection accuracy mAP of 77.6% in scenes with impurities and 99.3% across all scenes (Ban et al. 2023). Moreover, YOLOv8-HD demonstrated reduced memory size (Ban et al. 2023), GFLOPs, and faster inference time compared to YOLOv8, making it a valuable technical support tool for seed counting instruments in various scenarios.

Furthermore, ”Wheat Teacher,” developed by researchers from the College of Information and Intelligence at Hunan Agricultural University in Changsha, showcases the potential of semi-supervised learning in effectively utilizing unlabeled data to enhance model performance. By dynamically allocating pseudo-labels and filtering losses, the model demonstrates robustness and adaptability, particularly in scenarios where labeled data is scarce or costly to obtain. ”Wheat Teacher” employs a combination of two semi-supervised techniques: pseudo-labeling and consistency regularization. Noteworthy innovations include the Pseudo-label Dynamic Allocator, a dynamic threshold component tailored for wheat head detection scenarios, and the Loss Dynamic Threshold, which filters losses. Impressively, on the GWHD2021 dataset, ”Wheat Teacher” achieved a mean average precision (mAP@0.5) of 92.8% using only 20% labeled data, outperforming two fully supervised object detection models trained with 100% labeled data (Zhang et al. 2024). This underscores the significant improvements in mAP0.5 exhibited by the semi-supervised approach across various labeled data usage ratios.

In parallel to these indi1vidual research endeavors, collaborative initiatives like the Global Wheat Head Detection challenges, spanning 2020 and 2021, provided fertile ground for collective innovation in wheat head detection. These challenges, hosted on prominent platforms like Kaggle and AIcrowd, attracted researchers worldwide to showcase their solutions. Notably, the top-performing entries in both challenges consistently leveraged well-established open-source architectures, including EfficientDet, Faster R-CNN, YOLOv5, and YOLOv3. This collaborative spirit underscored the power of collective intelligence in tackling complex agricultural problems. The summary of the winning solutions is as shown in table 1.

A standout contribution emerged from Gong et al. in 2021, offering a fresh perspective on wheat head detection using deep neural networks (Gong et al. 2020). Their innovative methodology, documented in their paper, enhanced the backbone network with Spatial Pyramid Pooling (SPP) networks and employed sophisticated feature fusion strategies. The result was a significant leap in performance, with a mean average precision of 94.5%. Gong et al.’s rigorous evaluation, comparing their model against seven other methodologies, highlighted the robustness and versatility of their approach across diverse experimental conditions.

In the existing literature of wheat head detection methods, we examined three prominent approaches: Faster R-CNN, YOLO, and DETR. Faster R-CNN, renowned for its accurate object localization, employs a multi-stage architecture with region proposal networks and region-based CNNs. YOLO, on the other hand, is recognized for its real-time performance, utilizing a single-stage design and grid-based prediction mechanism. DETR represents a novel approach, leveraging transformer architecture to predict object class labels and bounding box coordinates directly. While each approach exhibits distinct strengths in detection performance, computational efficiency, and model complexity, we chose to evaluate DETR despite its potentially longer training times due to its transformer-based design. Faster R-CNN achieves high accuracy in wheat head detection but can be computationally intensive due to its multi-stage architecture. Meanwhile, YOLO achieves impressive detection speed but may struggle with small object detection and overlapping instances, which are common challenges in wheat head detection scenarios. DETR, on the other hand, offers a streamlined end-to-end architecture with remarkable accuracy, making it a compelling candidate for wheat head detection tasks that prioritize precise localization and robust performance in diverse conditions. By exploring DETR alongside Faster R-CNN and YOLO, we aim to provide a comprehensive understanding of the trade-offs and capabilities of different approaches in addressing the challenges of wheat head detection.

Preprocessing & Augmentation

The utilization of the Global Wheat Head Detection (GWHD) dataset 2020 by Zhang et al. (Zhang et al. 2022) marked a significant stride in wheat head detection research. Their study embraced over 3,000 images for training and around 1,000 for testing, encompassing a wide spectrum of conditions such as weather, illumination, and growth stages. Their preprocessing endeavor unfolded in two pivotal phases, each designed to fortify the dataset and address inherent challenges.

In the initial phase of **Data-set Analysis**, Zhang et al. meticulously scrutinized the characteristics of the dataset. They observed a distribution of detection frames within the training set, with the majority of images containing 20 to 60 detection boxes, while some lacked any, reaching a maximum of 116. The challenges posed by overlapping plants, blurred imagery, and phenotype variations were duly noted. Subsequently, in the **Data Augmentation** stage, a multifaceted approach was adopted. Basic augmentation techniques involved segmenting each original image into five sub-images, horizontally and vertically flipping them, and introducing alterations in the HSV channel to generate 15 augmented images per original, thus expanding the dataset to 45,000 images. Experimental augmentation methods, namely Cutout, CutMix, and Mosaic techniques, were then applied to the training samples. Cutout selectively removed portions of the sample while retaining the original label, CutMix integrated random samples from the training set into specific regions of the image, and Mosaic utilized a mosaic of multiple pictures to enrich the background of detected objects. These comprehensive preprocessing measures were tailored to heighten the model’s resilience against the myriad challenges posed by varying image conditions and object detection scenarios.

Similarly, Fourati et al. (Fourati, Mseddi, and Attia 2021) leveraged the GWHD dataset 2020 in their pursuit of wheat head detection excellence. Their study uncovered a prevalent bounding box coverage of 20 to 40% across most images, coupled with significant brightness variations. In line with Zhang et al., their preprocessing journey unfolded across three distinct phases, each aimed at refining dataset quality and augmenting model robustness. The first phase, **Data**

Year	Domain Data Augmentation	Architecture	Ensemble Approach	Challenge Score
GWC, 2020	Mixup and custom mosaic	EfficientDet & Faster RCNN	Random subsampling	$mAP_{2021} = 0.690$
	Mixup and cutmix	EfficientDet & FasterRCNN	Random Subsampling	$mAP_{2021} = 0.688$
GWC, 2021	Mixup Mosaic	YoloV3 YoloV5	No Domain subsampling	$mAP_{2021} = 0.700$
	Mosaic and cutmix	YoloV5	No	$mAP_{2021} = 0.695$
	Cutmix	YoloV4	Yes	$mAP_{2021} = 0.695$

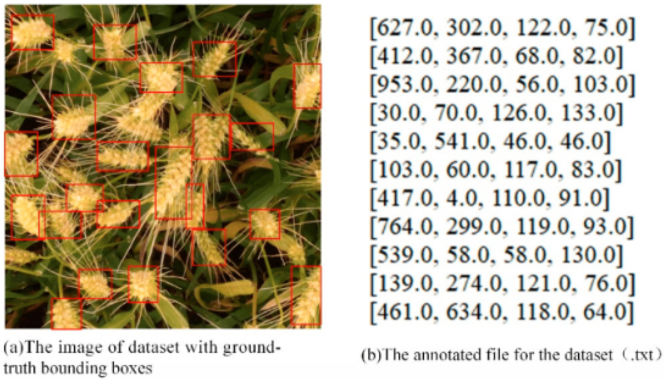
Table 1: Summary of the winning solutions.

Cleaning, saw Fourati et al. meticulously address issues pertaining to bounding boxes. They identified and removed excessively large and minuscule bounding boxes, presumed to be artifacts of labeling errors. Proceeding to the **Data Splitting** stage, they employed stratified k-fold splitting to ensure homogeneity in box distributions and source image counts across each fold. Finally, in the **Data Augmentation** phase, a suite of augmentation techniques including flips, rotations, cropping, and noise injections were judiciously applied, tailored to the specific context of wheat head detection. These concerted efforts were aimed at rectifying bounding box inconsistencies and elevating model generalization by augmenting dataset quality.

Likewise, the authors behind (Gong et al. 2020) set out on a path of data preprocessing, striving to elevate the efficiency of their wheat head detection model. They undertook measures such as removing images based on bounding box size from the original GWHD dataset. Additionally, a series of data enhancement operations including rotation, cropping, and noise addition were performed. These efforts underscore the shared objective across studies: to refine dataset quality and bolster model performance through comprehensive preprocessing strategies.

Dataset & Preprocessing

In this study, the Global Wheat Head Detection Dataset 2021 served as the primary data source. This dataset comprises over 6000 images, each with dimensions of 1024x1024 pixels, containing more than 300,000 distinct wheat heads, accompanied by their respective bounding boxes (David et al. 2021). Originating from 11 different countries and spanning across 44 distinct measurement sessions, this dataset offers a diverse representation of wheat cultivation scenarios worldwide. To provide a visual representation, a segment of the dataset showcasing the ground-truth bounding boxes and label files is depicted in Figure 1. The images from the dataset cover various environmental conditions, wheat varieties, and growth stages.



[627.0, 302.0, 122.0, 75.0]
[412.0, 367.0, 68.0, 82.0]
[953.0, 220.0, 56.0, 103.0]
[30.0, 70.0, 126.0, 133.0]
[35.0, 541.0, 46.0, 46.0]
[103.0, 60.0, 117.0, 83.0]
[417.0, 4.0, 110.0, 91.0]
[764.0, 299.0, 119.0, 93.0]
[539.0, 58.0, 58.0, 130.0]
[139.0, 274.0, 121.0, 76.0]
[461.0, 634.0, 118.0, 64.0]

(a)The image of dataset with ground-truth bounding boxes

(b)The annotated file for the dataset（.txt）

Figure 1: A segment of Global Wheat Head Detection Dataset

In preparing the data for model training, it was imperative to ensure compatibility with DETR's unique architecture. Unlike traditional object detection methods that rely on complex networks or predefined anchor boxes, DETR simplifies the process by using a transformer encoder-decoder design. This design enables the model to predict object bounding boxes and their associated class labels in a single pass, without the need for additional layers. As a result, the data needed to adhere to the COCO (Common Objects in Context) JSON format, a requirement imposed by DETR. This format includes detailed annotations for each image, specifying the precise location of objects within bounding boxes along with their corresponding class labels. Unlike CSV (Comma-Separated Values) format, which lacks the structured annotation capabilities necessary for object detection tasks, COCO JSON provides a standardized and comprehensive way to represent datasets.

Before images are input into the DETR model for training, they must first undergo preprocessing, which is made easier by the `DetrImageProcessor`. Resizing the photos to fit specific dimensions is a crucial part of this preprocessing. Each image is automatically resized by the `DetrImageProcessor` to have a minimum size of 800 pixels and a maximum size of 1333 pixels. The selection of these dimensions ensures consistency between the training and inference stages by matching them with the default values that DETR use during inference. However, the resizing process might put the GPU's memory to the test, particularly if it is done on a large number of photos. This is due to the GPU memory resources being used by each image after it has been flattened and processed by the convolutional layers.

Approach and Implementation

The DETR architecture, tailored specifically for end-to-end object detection using transformers, undergoes a meticulous training process to proficiently recognize the distinctive attributes of wheat heads. This entails considering factors such as their small size, diverse appearances, and potential occlusions. During training, batches of preprocessed data are fed into the model, iteratively adjusting its weights to minimize a predefined loss function. To ensure optimal performance and prevent overfitting, a validation strategy is implemented, monitoring the model's efficacy on a separate validation dataset. Periodic checkpoints are saved to track the model's progress, while crucial training metrics like loss, accuracy, and validation scores are logged for thorough analysis and refinement.

Throughout the training phase, several training parameters play a pivotal role in optimizing the model's performance. In this study, a batch size of 4 is utilized, meaning that during each training iteration, the model processes and updates its weights based on four examples simultaneously. Furthermore, the training dataset comprises 3657 examples, encompassing a diverse set of images and corresponding annotations. This diversity ensures that the model learns robust features for wheat head detection across various environmental conditions, wheat varieties, and growth stages, enhancing its adaptability and accuracy in real-world scenarios.

Parameter	Value
lr	1e-4
lr_backbone	1e-5
weight_decay	1e-4

Table 2: Model Parameters

Table further outlines the parameters used in the model, along with their corresponding values. Firstly, the ”lr” parameter, short for learning rate, is set to 1e-4. The learning rate determines the step size at which the model adjusts its weights during training based on the gradient of the loss function. A higher learning rate allows for faster convergence but risks overshooting the optimal solution, while a lower learning rate may lead to slower convergence but potentially more accurate results. Secondly, the ”lr_backbone” parameter denotes the learning rate specific to the backbone of the model, such as ResNet-50 in this context, and is set to 1e-5. The backbone network is responsible for feature extraction from input images and plays a crucial role in the overall performance of the model. Adjusting the learning rate for the backbone separately from the rest of the model allows for finer control over its training dynamics. Lastly, the ”weight_decay” parameter, set to 1e-4, controls the amount of regularization applied to the model's weights during training. Regularization helps prevent overfitting by penalizing large weight values, thus encouraging the model to generalize better to unseen data. A higher weight decay value increases the regularization strength, while a lower value may lead to potential overfitting.

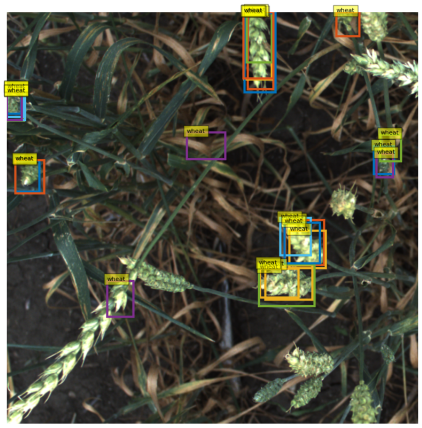


Figure 2: Result on one image of validation set on 3 epochs

Figure 2 shows the bounding box prediction of the model after training for only 3 epochs.

Evaluation

The performance of the model is evaluated using Mean Average Precision (mAP), a widely adopted metric for object detection tasks. mAP offers a comprehensive assessment by considering both the model's ability to correctly identify objects (precision) and its capability to detect all instances of a particular class (recall).

To calculate mAP, we rely on two fundamental concepts: Intersection over Union (IoU) and Precision-Recall.

Intersection over Union (IoU)

IoU quantifies the overlap between two bounding boxes. It's calculated as the ratio of the area shared by the predicted bounding box and the ground truth bounding box divided by their combined total area.

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

A higher IoU value signifies a greater degree of overlap, indicating a more accurate prediction. IoU often serves as a threshold for classifying a prediction as a true positive.

Precision & Recall

Precision measures the proportion of correctly predicted positive samples (True Positives, TP) amongst all predicted positive samples (TP + False Positives, FP).

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

Recall (also known as sensitivity) assesses the model's ability to detect all actual positive instances. It's the ratio of correctly identified positive samples (TP) to the total number of actual positive samples (TP + False Negatives, FN).

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

Here, TP represents correctly predicted positive samples, FP indicates incorrectly classified positive samples (model predicted positive, but actually negative), and FN denotes missed positive samples (model predicted negative, but actually positive).

mean Average Precision (mAP)

mAP is the average of the Average Precision (AP) scores across all object categories present in the dataset. AP for each class is typically determined using a Precision-Recall curve. This curve depicts the trade-off between precision and recall for various IoU thresholds. The area under the Precision-Recall curve (AUC) serves as the AP for that specific class.

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N}$$

where,

- N denotes the total number of object categories
- AP_i represents the Average Precision for category i.

Experimental Results

The investigation into the fine-tuned DETR model with a ResNet-50 backbone for wheat head detection yielded promising results. The evaluation of the model's performance utilized the mAP metric at two distinct IoU thresholds: 0.5 and 0.75, offering detailed insights into the model's proficiency in detecting and precisely localizing wheat heads across varying degrees of overlap.

One of the pivotal findings was the significant impact of training epochs on the model's performance. When trained for a relatively modest duration of 5 epochs, the model exhibited a commendable mAP@50 score of 0.417, indicative of a moderate success rate in identifying wheat heads with satisfactory overlap between predicted bounding boxes and ground truth annotations. This underscores the model's adaptability to diverse wheat head appearances across a spectrum of images. However, the observed lower mAP@75 value of 0.224 highlights a challenge in achieving highly precise localization, particularly when demanding a 75% overlap between predicted and ground truth bounding boxes.

Extending the training regime to 10 epochs resulted in a marked enhancement in performance. The model's mAP@50 increased impressively to 0.914, indicating a substantial improvement in overall detection accuracy. Notably, the mAP@75 also exhibited a significant leap to 0.729, signaling a heightened proficiency in precisely localizing wheat heads with stringent overlap criteria. This suggests that the model substantially benefits from prolonged training iterations, enabling it to discern subtle variations and accurately pinpoint wheat heads amidst the complexities inherent in field environments.

These results underscore the inherent potential of the DETR model for wheat head detection tasks. Not only does it achieve commendable overall accuracy (mAP@50), but it also demonstrates a promising capability for precise localization (mAP@75) with adequate training. Nevertheless, the discernible gap between mAP@50 and mAP@75 highlights the persistent challenge of achieving perfect localization for all wheat heads, particularly when considering factors such as varying object sizes and potential occlusions in real-world field settings.

The results show that our finetuned models outperforms some of the models proposed in recent literature, illustrated in Table 3. The best performing YOLO like singe stage model of (Zhang et al. 2022) achieved only an mAP score of 0.6893 with an input resolution of 512 × 512. In addition to this our model also outperforms the three winning solutions of the global wheat head detection challenges 2020 and 2021 since the results shown in 1 were obtained with an IoU threshold of 0.75 and our results exceed these scores.

However, Our model was not able to surpass the three part network proposed by (Gong et al. 2020) which was able to achieve a mAP score of 0.945 on an IoU threshold of 0.5 and a score of 0.545 on an IoU threshold of 0.95.

This investigation not only provides valuable insights into the performance nuances of the DETR model but also sets the stage for further exploration and optimization. Leveraging sophisticated data augmentation techniques to introduce diverse variations in the training data could bolster the model's resilience to real-world complexities and enhance its generalization capabilities. Additionally, meticulous fine-tuning of hyperparameters holds promise for further optimizing the model's performance, particularly concerning wheat head detection tasks.

Model	**Epochs**	**mAP@50**	**mAP@75**
facebook/detr-resnet-50	5	0.417	0.224
facebook/detr-resnet-50	10	0.914	0.729

Table 4: mAP Metric Results

Conclusion and Future Work

This study showcases the potential of the DETR model with a ResNet-50 backbone, for wheat head detection, offering promising results in terms of overall accuracy and precise localization. Evaluating the model's performance across various training epochs revealed significant improvements in detection accuracy, particularly with extended training durations. However, challenges remain evident, as indicated by the disparity between mean Average Precision values at different Intersection over Union thresholds, highlighting the complexity of achieving perfect localization under diverse field conditions. Future research endeavors could explore avenues for enhancing model robustness through advanced data augmentation techniques and fine-tuning of hyperparameters. Additionally, extending the study to encompass a broader range of environmental factors and wheat varieties could provide deeper insights into the model's performance and further refine its capabilities for real-world applications in agricultural settings. This research lays the groundwork for continued advancements in automated wheat phenotyping, contributing to the sustainable improvement of crop yield and global food security. The research could also be

Paper	Architecture	Dataset	mAP@0.5	mAP@0.75	mAP@0.95	mAP@0.5:0.95	mAP
Zhang et al. 2022	Based on YOLOv4	GWHD 2020	-	-	-	-	0.6893
Zang et al. 2022	Improved YOLOv5	GWHD 2021	0.951	-	-	0.545	-
Gong et al. 2020	Three Part YOLOv4	GWHD 2020	0.945	-	0.545	-	-
David et al. 2023	EfficientDet and Faster-RCNN	GWHD 2020	-	0.474	-	-	-
David et al. 2023	EfficientDet	GWHD 2020	-	0.690	-	-	-
David et al. 2023	YOLOv3	GWHD 2020	-	0.688	-	-	-
David et al. 2023	YOLOv5	GWHD 2021	-	0.700	-	-	-
David et al. 2023	YOLOv5	GWHD 2021	-	0.695	-	-	-
David et al. 2023	YOLOv4	GWHD 2021	-	0.695	-	-	-
Meng et al. 2023	YOLOv7-MA	GWHD 2021	-	-	-	-	0.9386
Ban et al. 2023	YOLOv8-HD	GWHD 2021	-	-	-	0.582	0.776
Zhu et al. 2022	FR-Transformer	GWHD 2021	0.883	0.385	-	-	-
Zhang et al. 2024	Semi-supervised learning	GWHD 2021	0.928	-	-	-	-

Table 3: Results in Previous Wheat Head Detection Approaches

extended by incorporating more recent variations of DETR such as Real Time DETR that claims to have surpassed YOLO architectures in detection and efficiency.

References

Ban, X.; Liu, P.; Xu, L.; and Zhao, J. 2023. A lightweight model based on yolov8n in wheat spike detection. 1–6.

Batin, M.; Islam, M.; Hasan, M. M.; Azad, A.; Alyami, S. A.; Hossain, M. A.; and Miklavcic, S. J. 2023. Wheatspikenet: an improved wheat spike segmentation model for accurate estimation from field imaging. *Frontiers in Plant Science* 14:1226190.

Bi, K.; Jiang, P.; Li, L.; Shi, B.; and Wang, C. 2010. Non-destructive measurement of wheat spike characteristics based on morphological image processing. 26:212–216.

David, E.; Serouart, M.; Smith, D.; Madec, S.; Velumani, K.; Liu, S.; Wang, X.; Espinosa, F. P.; Shafiee, S.; Tahir, I. S. A.; Tsujimoto, H.; Nasuda, S.; Zheng, B.; Kichgessner, N.; Aasen, H.; Hund, A.; Sadhegi-Tehran, P.; Nagasawa, K.; Ishikawa, G.; Dandrifosse, S.; Carlier, A.; Mercatoris, B.; Kuroki, K.; Wang, H.; Ishii, M.; Badhon, M. A.; Pozniak, C.; LeBauer, D. S.; Lilimo, M.; Poland, J.; Chapman, S.; de Solan, B.; Baret, F.; Stavness, I.; and Guo, W. 2021. Global wheat head dataset 2021: more diversity to improve the benchmarking of wheat head localization methods.

El-Hendawy, S.; Al-Suhaibani, N.; Mubushar, M.; Tahir, M. U.; Marey, S.; Refay, Y.; and Tola, E. 2022. Combining hyperspectral reflectance and multivariate regression models to estimate plant biomass of advanced spring wheat lines in diverse phenological stages under salinity conditions. *Applied Sciences* 12(4):1983.

Fourati, F.; Mseddi, W. S.; and Attia, R. 2021. Wheat head detection using deep, semi-supervised and ensemble learning. *Canadian Journal of Remote Sensing* 47:198–208. Received 19 Sep 2020, Accepted 13 Mar 2021, Published online: 29 Apr 2021.

Gong, B.; Ergu, D.; Cai, Y.; and Ma, B. 2020. A method for wheat head detection based on yolov4.

Meng, X.; Li, C.; Li, J.; Li, X.; Guo, F.; and Xiao, Z. 2023. Yolov7-ma: Improved yolov7-based wheat head detection and counting. *Remote Sensing* 15.

Mia, J.; Bijoy, H. I.; Uddin, S.; and Raza, D. M. 2021. Real-time herb leaves localization and classification using yolo. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–7.

TANG, L.; GAO, H.; Yoshihiro, H.; Koki, H.; Tetsuya, N.; sheng LIU, T.; Tatsuhiko, S.; and jin XU, Z. 2017. Erect panicle super rice varieties enhance yield by harvest index advantages in high nitrogen and density conditions. *Journal of Integrative Agriculture* 16(7):1467–1473.

Tian, H.; Wang, T.; Liu, Y.; Qiao, X.; and Li, Y. 2020. Computer vision technology in agricultural automation—a review. *Information Processing in Agriculture* 7(1):1–19.

Velu, G., and Singh, R. P. 2013. Phenotyping in wheat breeding. *Phenotyping for plant breeding: applications of phenotyping methods for crop improvement* 41–71.

Zhang, Y.; Li, M.; Ma, X.; Wu, X.; and Wang, Y. 2022. High-precision wheat head detection model based on one-stage network and gan model. *Frontiers in Plant Science* 13:787852. Original research article, Technical Advances in Plant Science, Published on 02 June 2022, Research Topic: Digital Innovations in Sustainable Agri-Food Systems.

Zhang, R.; Yao, M.; Qiu, Z.; Zhang, L.; Li, W.; and Shen, Y. 2024. Wheat teacher: A one-stage anchor-based semi-supervised wheat head detector utilizing pseudo-labeling and consistency regularization methods. *Agriculture* 14(2).

Zhu, Y.; Cao, Z.; Lu, H.; Li, Y.; and Xiao, Y. 2016. In-field automatic observation of wheat heading stage using computer vision. *Biosystems Engineering* 143:28–41.

Zhu, J.; Yang, G.; Feng, X.; Li, X.; Fang, H.; Zhang, J.; Bai, X.; Tao, M.; and He, Y. 2022. Detecting wheat heads from uav low-altitude remote sensing images using deep learning based on transformer. *Remote Sensing* 14(20).