Revolutionizing Rice Grain Counting: An Innovative TCLE–YOLO Model for Accurate Rice Grain Detection and Counting in Agricultural Yield Estimation

Faiq Shah[1*], Kamran Shah[1], Izhar Ul Haq[1]

[*] Department of Mechatronics Engineering University of Engineering & Technology Peshawar, Pakistan

Abstract- Precise measurement of the thousand-grain weight is critical for accurately forecasting rice yields. This parameter is essential for variety development and appropriate cultivation management. Precise identification and enumeration of rice grains are required for accurate measurements of thousand-grain weight, a pivotal stage in the research process. However, the procedure has substantial challenges due to the small size of rice grains, their intrinsic similarity, and the differing degrees of stickiness. The TCLE-YOLO model is an advanced deep learning technique that integrates a transformer encoder and coordinate attention module. It utilises the robust YOLOv5 as the underlying network architecture. The model integrates a coordinate attention (CA) module into the YOLOv5 backbone to enhance feature representation in small target regions. The system also incorporates a specialist detection head designed for detecting small targets. This head utilises a feature map with high resolution and low-level details. Furthermore, the neck module employs a transformer encoder to augment the network's capacity to include a broader spectrum of data and amplify the extraction of crucial characteristics from recognised targets. This increases the sensitivity of the additional detecting head specifically towards rice grains, particularly those that have a significant level of adhesion. The implementation of EIoU loss significantly improves accuracy. The results of our experiments on our custom rice grain dataset demonstrate outstanding precision, recall, and mAP@0.5 scores of 99.20%, 99.10%, and 99.20%, respectively, surpassing several state-of-the-art models. The TCLE-YOLO model we have developed offers a robust foundation for accurately identifying and quantifying rice grains. It provides vital information for accurate measurements of thousand-grain weight and the efficient assessment of rice breeding procedures.

Index Terms- About four key words or phrases in alphabetical order, separated by commas. Keywords are used to retrieve documents in an information system such as an online journal or a search engine. (Mention 4-5 keywords)

I. Introduction

To obtain accurate estimates of rice yields, it is essential to precisely measure the thousand-grain weight. This parameter is critical for variety breeding and efficient crop management. Accurate identification and enumeration of individual rice grains are essential for obtaining precise measurements of thousand-grain weight, a critical stage in the research process. Nevertheless, the method encounters substantial obstacles as a result of the diminutive dimensions of rice grains, their indistinguishable appearance, and the fluctuating degrees of adhesiveness. The TCLE-YOLO model is a cutting-edge deep learning method that integrates a transformer encoder and coordinate attention module, leveraging the powerful YOLOv5 as the underlying network architecture. In order to improve the representation of features in small target regions, the model smoothly incorporates a coordinate attention (CA) module into the YOLOv5 backbone. In addition, it includes a dedicated detection head that is specifically designed to identify small targets. This is achieved by utilising a feature map that has a high level of resolution and includes fine-grained details. In addition, the neck module integrates a transformer encoder to enhance the network's ability to capture a wider range of information and enhance the extraction of important properties from selected targets. This enhances the responsiveness of the supplementary detection component to rice grains, especially those that have a substantial level of adhesion. The implementation of EIoU loss greatly enhances accuracy. The results of our experiments on our custom rice grain dataset show exceptional precision, recall, and mAP@0.5 scores of 99.20%, 99.10%, and 99.20%, respectively. These scores surpass those achieved by other cutting-edge models. The TCLE-YOLO model demonstrates exceptional efficacy in the identification and quantification of rice grains. This study offers useful insights that may be used to accurately quantify the thousand-grain weight and effectively evaluate rice breeding procedures.

Deep learning-based item detection approaches provide accurate and reliable crop counting, in contrast to the previously described image processing methods. This is due to their exceptional ability to extract strong features and their capacity for autonomous learning [8,9,10]. Khaki et al. [11] conducted a study where they proposed a sliding window strategy that employed a CNN classifier to precisely identify and quantify maize cob grains. Their approach produced remarkable outcomes, attaining an RMSE of 8.16% while calculating the mean quantity of grains in a grain-counting assignment. Gong et al. [12] conducted a study where they created a fully convolutional network that accurately identified grains within a panicle. The network achieved an excellent accuracy rate of 95%. Tan et al. [13] employed YOLOv4 to detect cotton seedlings in individual frames in their study. In addition, they utilised an optical flow-based tracking technique to estimate the movements of the camera and ascertain the number of cotton seeds. Lyu et al. [14] did a study where they effectively detected and measured green citrus fruits in orchards using an improved version of YOLOv5. Their methodology entailed integrating a convolutional block attention module with a detection layer. The experiment's findings showcased that the suggested model attained a remarkable mean average precision (mAP) of 98.23% at a threshold of 0.5, along with a recall rate of 97.66% specifically for green citrus fruits. Rong et al. [15] devised a tomato cluster identification technique by utilising an improved YOLOv5-4D. By utilising both RGB photos and depth images as input, they achieved a remarkable accuracy of 97.9% and a mean average precision (mAP) of 0.748 at the intersection over union (IoU) threshold range of 0.5 to 0.95.

Deep learning-based object detection algorithms provide a highly accurate and efficient approach for counting targets. However, due to the small size of rice grains, their remarkably similar appearances, and the lack of intricate details in each grain, the visual characteristics that can successfully differentiate distinct sticky grains are usually limited to a small, specific area. As a result, these characteristics are not easily assimilated by a network. Consequently, the detector may fail to detect all the grains, resulting in a reduction in counting accuracy and ultimately impacting the precision of yield calculations. Furthermore, there is a widespread acknowledgment that the weight of a thousand grains is closely associated with many features of the grains, such as their length, width, thickness, and the ratio of kernel length to width [16]. Precise identification and enumeration of rice grains is essential for getting accurate grain-size characteristics utilising imaging technology. Additionally, it plays a crucial role in assessing farming techniques, examining seed characteristics, developing new strains, and efficiently classifying rice kernels. Therefore, to effectively identify and quantify rice grains, it is essential to extract and exploit significant feature data from specific regions of the grains. Deep learning algorithms are crucial in this process, especially for accurately identifying highly sticky rice grains that provide difficulties in distinguishing.

According to our comprehension, the attention module exhibits resemblances to the human visual attention mechanism. It prioritises specific regional information while discarding secondary input, similar to how the human brain processes tasks. This attribute significantly enhances the effectiveness of the model in processing input and can have a pivotal impact on improving the feature perception of network models. Additionally, this technology can be used to detect and locate minuscule objects that may be concealed or adhered together [17]. For instance, Peng et al. [18] successfully developed a soybean aphid recognition model by utilising a Convolutional Neural Network (CNN) with an attention mechanism, resulting in enhanced accuracy. Zhang et al. [19] did a study where they integrated an attention mechanism and a residual network (ResNet) into a system for identifying flawed wheat grains. The findings demonstrated a notable enhancement in the precision of categorising flawless and different forms of flawed wheat grains. The significance of including an attention mechanism into object detection algorithms was emphasised in our study. This upgrade has been empirically demonstrated to significantly enhance the precision of the algorithms, rendering it one of the most efficacious improvements among a multitude of modifications.

Due to the constraints of current image analysis technologies in detecting and counting grains, as well as the challenges posed by sticky rice grains, it is necessary to create a deep learning model that can effectively meet these particular demands. The model should possess robust feature recognition abilities, prioritising the accurate identification of features in the designated areas to ensure precise adhesion recognition. The system should possess the capability to distinguish between cohesive rice grains and autonomously identify specific areas with varying degrees of stickiness. Furthermore, the model should possess the capability to autonomously identify and pinpoint minute targets that exhibit highly similar attributes, hence augmenting its proficiency in detecting rice grains.

This work endeavours to design a technique for detecting and quantifying rice grains through thorough investigation and analysis. The suggested methodology integrates an attention mechanism with YOLOv5, denoted as TCLE-YOLO. To address the problem of distinguishing sticky rice grains, we integrate the coordinate attention (CA) module [21] into the YOLOv5 backbone module. The purpose of this integration is to augment the model's capacity to focus on minute objectives, hence enhancing the network's proficiency in expressing characteristics. In addition, a dedicated detecting head has been created to improve the identification of small targets. This head is specifically engineered to possess exceptional sensitivity and is founded upon a transformer encoder's high-resolution feature map. The objective of this modification is to strengthen the overall capacity for recognising diminutive targets [22]. In addition, the transformer encoder is employed in the neck module to expand the network's receptive area and prioritise the important feature information related to the rice grain region. This increases the responsiveness of the supplementary detection component to smaller entities. The subsequent sections of this document are structured in the following manner: Section 2 provides a detailed description of the data sources and research methodologies employed in the creation of TCLE–YOLO. Section 3 focuses on the analysis of the experimental findings and facilitates in-depth conversations. Section 4, ultimately, provides the results derived from our investigation.

II. Materials & Methods

A. Pre-processing & Image Acquisition

The experimental materials utilised in this investigation were cultivated at the Rice Research Institute, Anhui Academy of Agricultural Sciences. The rice grains were meticulously arranged on a spotless white background board in anticipation of capturing the photograph. The photos were acquired via a camera, as depicted in Figure 1, with a mobile phone camera affixed to a visual platform. The camera holder's vertical adjustment facilitated the replication of various distances. The investigation involved capturing photos of rice grains with different levels of adhesion using a cell phone camera. The distances from the white background board were 8 cm, 15 cm, and 20 cm. According to the criteria outlined in [23], a local picture area containing 210 grain adhesions was categorised as having 'mild adhesion,' while 1020 grains were categorised as 'moderate adhesion,' and more than 20 grains were branded as 'severe adhesion.' There were 1000 rice grain photographs recorded, each having a resolution of 3024 × 4032. The photos comprised of 250 rice grains with mild adhesion, 400 rice grains with moderate adhesion, and 350 rice grains with severe adhesion.

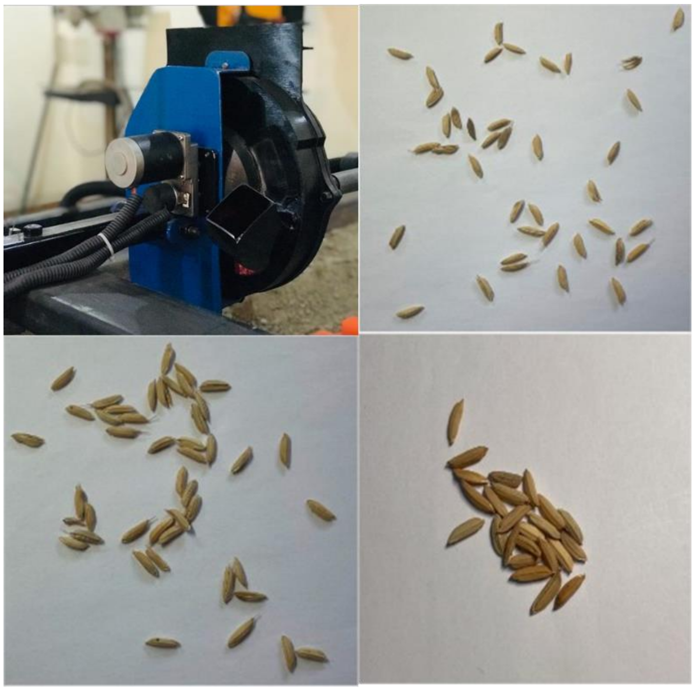


Figure 1. Seed Planter Image Acquisition and Dataset Samples

Figure 1 displays a collection of sample photos illustrating rice grains with different degrees of adherence. The photographs were acquired in RGB format, with the original images having a far greater resolution than the required input image size for YOLOv5. Managing the high resolution of the original photos may provide a difficulty during network training. It could potentially exert excessive pressure on the GPU memory, resulting in training failures. To resolve this issue, the photos of rice grains that were taken were reduced to a resolution of 640 × 640. The resolution change was implemented to strike a balance between maintaining crucial image features and improving model training for optimal efficiency and success, taking into account resource constraints.

B. Annotation & Image Augmentation

In order to successfully train a deep neural network, it is important to possess a substantial quantity of images. One established approach to enhance the performance of the model is to expand the dataset through the utilisation of diverse methodologies [24]. The utilisation of mosaic data augmentation, as depicted in Figure 2, entails the stochastic selection of four photos. Subsequently, these photos are proportionally adjusted, mirrored, and systematically organised to create a composite image. This method facilitates the process of recombination and amplifies the dataset [25]. Furthermore, diverse data augmentation techniques were utilised to improve the original photos, leading to a significant expansion in the dataset size for rice grain detection and counting. The techniques encompassed brightness conversion, multi-angle image rotation, and noise addition. Upon implementing data augmentation techniques, the dataset's sample count was augmented to 6000.

Ground truth (GT) images for object identification model training were labelled using bounding boxes. The labelling technique employed the LabelImg annotation software for accurate hand annotation. The regions containing individual rice grains were delineated using green bounding boxes. The annotation results were saved in XML files and used to train the model on the training dataset. Furthermore, they were employed to assess the model's performance on the validation and test datasets. The rice grain photos were partitioned into training and validation datasets at a ratio of 7:2, with the remaining images serving as the test set to assess the performance and resilience of the proposed model.

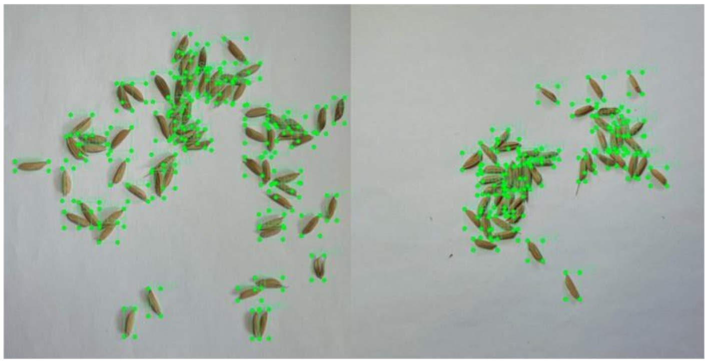


Figure 2. Annotated Sample images (Moderately & Severely adhered)

C. Rice Grains Detection and Counting Model

Experienced researchers are undoubtedly acquainted with the YOLO (You Only Look Once) framework, known for its efficient single-stage detection abilities. Throughout its development, the framework has seen multiple changes, ultimately resulting in the emergence of YOLOv5 as a particularly noteworthy version. YOLOv5 is renowned for its concise architecture, versatility, and remarkable efficiency in analysing images, rendering it a superb option for real-time object detection models [26]. The YOLOv5 comprises three main components: the backbone, neck, and head. The backbone module plays a vital role in YOLOv5 by extracting features from input photos. The neck module effectively combines information from several network tiers by up-sampling features from the backbone module via both bottom-up and top-down pathways. The detection head module in target detection jobs is responsible for anticipating picture attributes and overseeing category, position, confidence, and other pertinent elements. This article utilised the YOLOv5 framework as the foundation for the specified detection and counting methodology. However, the inclusion of numerous convolutions in the feature pyramid network of YOLOv5 may lead to information loss during training, particularly for small targets [27]. The precise identification and quantification of rice grains can be a difficult task due to their diminutive size, limited resolution, and comparable colour and structure, which frequently leads to their sticking together and obstructing one another. The detection of small targets requires the utilisation of high-resolution representations, which may not be adequately captured in deep-layer features.

In order to enhance the model's accuracy and adaptability in identifying and quantifying rice grains, we have integrated a Coordinate Attention (CA) module into the YOLOv5 backbone. In addition, we have incorporated a transformer encoder into the neck module. Furthermore, a detection head with a high level of sensitivity towards objects of small dimensions was integrated. Figure 3 displays the structure of our rice grain detecting model. Placing the coordinate attention module ahead of the SPFF block in the backbone module improves its ability to capture complex information from smaller targets, similar to that of an expert researcher. By employing the transformer encoder, the model efficiently enhances its ability to perceive a wider range of information and focuses its attention on important features within the region of the rice grain. The output of the transformer encoder is a feature map that has both low-level details and high-resolution. Subsequently, the input is directed towards the detecting head, which is explicitly engineered to identify and locate targets of reduced dimensions.

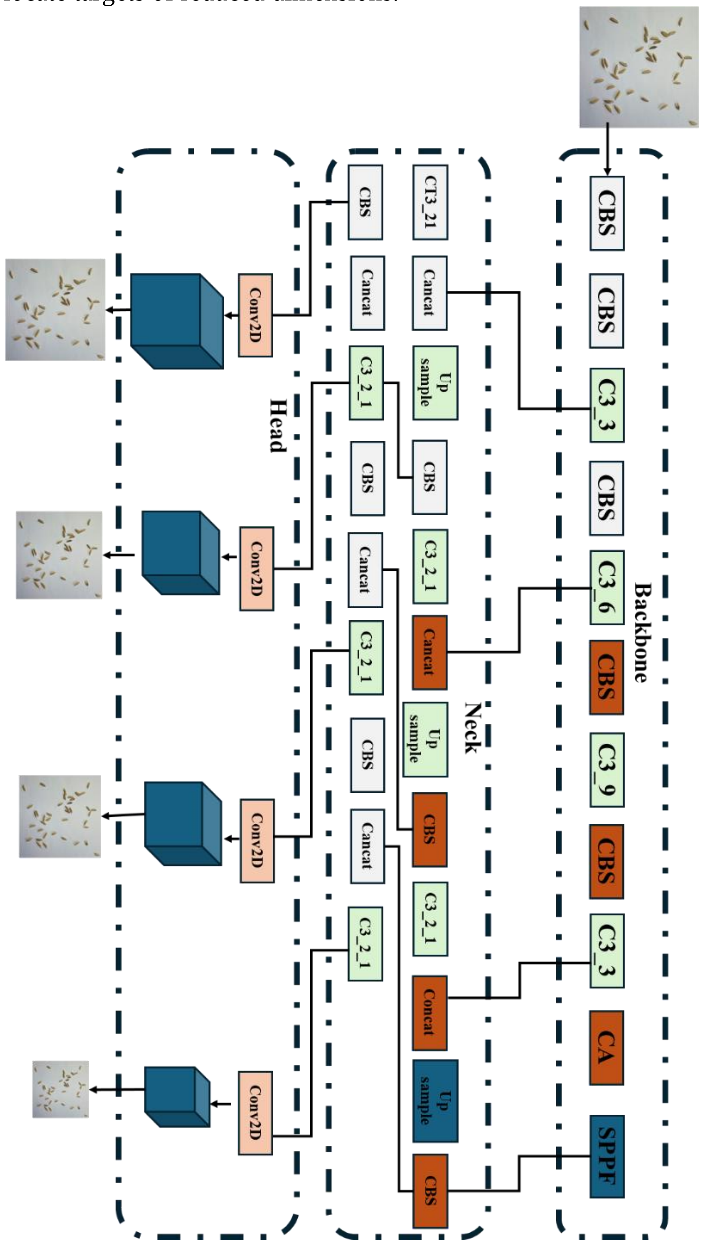


Figure 3. Developed Methodology

D. Enhanced Backbone Subsystem

The YOLOv5 model's backbone module is specifically engineered to effectively extract features from images. This is accomplished by employing a blend of CBS and C3_x blocks, which are subsequently merged together using an SSPF. However, the progression of visual characteristics from superficial to profound layers, coupled with a reduction in magnitude, can occasionally lead to the exclusion of significant semantic particulars, especially when addressing diminutive objects. Preserving semantic and location information is crucial for accurate recognition of rice grains, particularly in measuring thousand-grain weight, due to their small size, high overall similarity, and varied adhesion degrees. Expanding upon the discoveries made by Li et al. [28], which emphasised the significance of an attention mechanism in directing deep learning networks towards crucial characteristics, we integrated a Coordinate Attention (CA) module prior to the SSPF block in the foundational structure, as illustrated in Figure 4.

The CA module possesses the capability to incorporate coordinate information and produce attention maps. When dealing with a feature map X with dimensions of C × H × W, the coordinate information is embedded by encoding each channel along the horizontal and vertical coordinates. This is accomplished by employing two spatial pooling kernels: (H, 1) and (1, W). Two attention maps are produced, one for the horizontal direction with dimensions C × H × 1, and another for the vertical direction with dimensions C × 1 × W. The attention maps illustrate the existence of the region of interest in the associated row and column. They capture long-range connections in one spatial direction with precise positional details, similar to the approach of a skilled researcher. The two feature maps are merged in the channel dimension and subsequently undergo convolutional processing, yielding an intermediate feature map with dimensions of C × (W + H) × 1. To generate attention weight maps in both the horizontal and vertical directions, the intermediate feature map is partitioned along the spatial dimension. Subsequently, the feature maps obtained from the split are subjected to separate convolution operations. The feature maps undergo normalisation using the sigmoid function to produce attention weight maps for both the horizontal and vertical orientations. Multiplying the weight maps with the input X yields an augmented feature map. The utilisation of the CA module enhances the accuracy of the coordinate data for rice grains, hence generating feature maps that exhibit heightened sensitivity to both direction and position. This aids in the precise identification of rice grains by the detection model, particularly when they are clumped together.

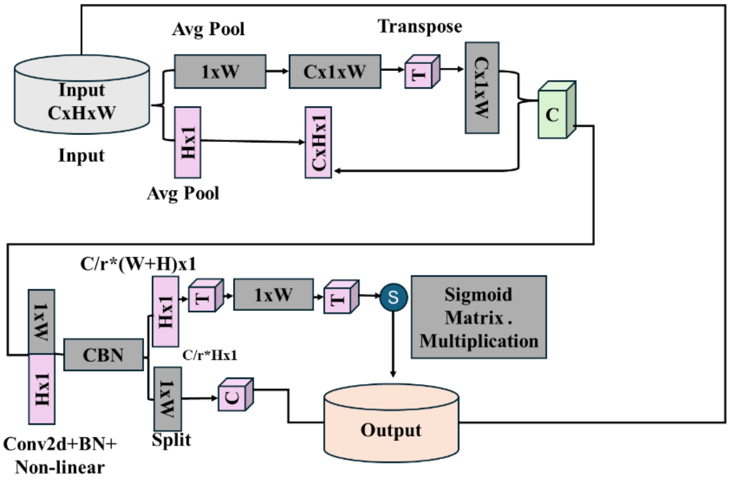


Figure 4. Co-ordinate attention module

E. System Architecture for Various Detection Heads

The detecting head module, positioned posterior to the neck module, is tasked with forecasting classifications and producing bounding boxes by analysing visual characteristics. When handling smaller targets, the feature representation may have restricted semantic information, which might pose challenges in precisely positioning bounding boxes in comparison to bigger targets. During the regression phase of the detection head, there is a small deviation of one pixel in the bounding box used for regression. The offset greatly affects the precision of forecasting small targets. According to reference [27], any divergence in the prediction bounding box can affect the network's ability to correctly classify the target during the classification phase of the detecting head. Convolutional neural networks frequently encounter difficulty in detecting little objects as a result of the restricted feature information included within the deep feature maps they assess. Conversely, features that possess intricate spatial information exhibit higher resolutions and excel at recognising small targets. To address this problem, a detection head with a high level of sensitivity to small targets was integrated. Figure 5 illustrates that the prediction head was obtained using the low-level, high-resolution feature map in order to enhance the ability to detect small targets. A neck module was incorporated into the YOLOv5 framework to augment the feature extraction process that occurs between the backbone and head modules. A transformer encoder block, which includes a self-attention mechanism, was smoothly integrated into the YOLOv5 C3 block. This improvement greatly enlarges the area of perception in the neck, allowing for accurate identification and pinpointing of even the most minute targets. The transformer encoder block, depicted in Figure 5, comprises two sub-layers. The first sub-layer efficiently distributes feature dimensions among several single-head self-attention mechanisms using a multi-head attention mechanism. This strategic approach emphasises the importance of specific areas within an image. The network can prioritise vital information and identify targets precisely by assessing numerous properties using advanced approaches. By merging various attention outcomes, this procedure allows the network to acquire significant contextual semantic data. The feed-forward neural network comprises a multi-layer perceptron (MLP) as a fully linked layer, which efficiently prevents feature degradation by utilising a second sub-layer. Residual connections are used to connect the fusion features of the two sub-layers, while layer normalisation is applied before and after both sub-layers to enhance convergence and reduce overfitting. By incorporating the transformer encoder prior to the detecting head for small objects, it enables a thorough representation of global information throughout the entire image. This facilitates the transfer of contextual information and ultimately enhances the identification of rice grains with different levels of adhesion, akin to that of an expert researcher.

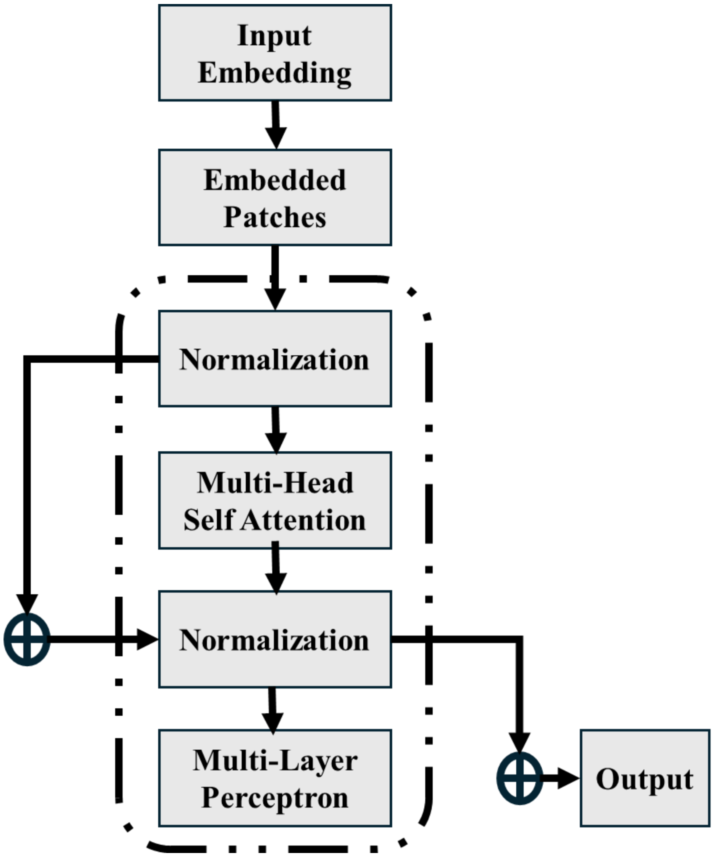


Figure 5. Transformer encoder block

III. Results

The experimental configuration utilised the Windows 10 operating system, executed with the PyTorch deep learning framework employing Torch version 1.13 and CUDA version 11.6. In my research, I employed state-of-the-art computing resources, including a high-performance NVIDIA GeForce RTX 4080 graphics card, an Intel I7-11700K CPU, and a substantial 32GB DDR4 3200 memory. The training method employed an input image resolution of 640 × 640, utilised SGD as the optimisation function, executed for 200 epochs, employed a batch size of 16, and initiated with a learning rate of 0.01. The rice grain dataset was utilised for both the training and testing phases of the programme.

To assess the model's detection ability, we employed precision (P), recall (R), and mean average precision (mAP) as the assessment metrics. The metrics are precisely described in the following manner:

$$\boldsymbol{Precision} = {}^{\boldsymbol{TP}}\!/_{\boldsymbol{TP+FP}}$$

$$\boldsymbol{Recall} = {}^{\boldsymbol{TP}}\!/_{\boldsymbol{TP+FN}}$$

$$\boldsymbol{AP} = \int_{\boldsymbol{0}}^{\boldsymbol{1}} \boldsymbol{P(R)dR}$$

$$\boldsymbol{mAP} = {}^{\sum \boldsymbol{Ap}}\!/_{\boldsymbol{N}}$$

Above equations define the terms "true positive" (TP) and "false positive" (FP) in the context of rice grain detection. A true positive refers to the right identification of rice grains, while a false positive refers to the mistaken identification of backgrounds as rice grains. A false negative (FN) refers to the incorrect identification of rice grains as background. Accurately determining the ground truth box and the prediction box produced by the detection model is essential in order to achieve true positives (TP) and false positives (FP). The prediction box contains crucial information, including the category, confidence score, and coordinate data. The forecast results are arranged in descending order based on their confidence levels, excluding any scores below 0.5. The Intersection over Union (IoU) values are computed to identify the maximum overlap between the predicted bounding box and the ground truth bounding box. If the intersection over union (IoU) value is more than 0.5 and it is the first match, it is classified as a true positive (TP). Alternatively, it is designated as FP. Consequently, a higher true positive (TP) value signifies a better probability of accurate prediction and superior model detection performance, whereas an increase in false positive (FP) suggests more incorrect detections and a decrease in model performance. In research, precision refers to the proportion of accurately identified results among all the detected results, while recall is the ratio of accurately identified findings to the total number of real results. The mean average precision (mAP) is a quantitative statistic that evaluates the overall detection efficacy of a model. The calculation involves computing the mean of the average precision (AP) for each category, which corresponds to the area under the precision-recall curve. This offers a more equitable and thorough evaluation of the model's overall performance. The mAP@0.5 metric denotes the average precision, calculated as the mean value, using an Intersection over Union (IoU) threshold of 0.5. Conversely, mAP@0.5:0.95 calculates the mean average precision (mAP) by considering several intersection over union (IoU) thresholds. These thresholds range from 0.5 to 0.95, increasing by 0.05 increments. During our research, we trained multiple detection models utilising the same datasets and experimental setups. Once the evaluation metrics on the validation set stopped improving after a certain training epoch, it became evident that the models had reached a point of convergence on the dataset. In order to evaluate their performance, the precision, recall, mAP@0.5, and mAP@0.5:0.95 of each trained model were compared on the test set.

A. Experiments for Ablation

As part of our investigation, we conducted ablation experiments to assess the efficacy of several modules in the TCLE–YOLO model, utilising YOLOv5 as the underlying network. Table 1 displays the precision, recall, mAP@0.5, and mAP@0.5:0.95 metrics achieved by different modules, including the coordinate attention module and the transformer encoder block stated previously. Extensive study and analysis have revealed that TCLE-YOLO demonstrates exceptional detecting skills for rice grains. This can be due to its distinctive amalgamation of a CA module, transformer encoder block, small target prediction head, and EIoU loss. It exhibited exceptional performance in terms of precision, recall, mAP@0.5, and mAP@0.5:0.95 metrics, surpassing other approaches.

When comparing TCLE–YOLO with YOLOv5, it is crucial to emphasise the enhancement in accuracy, which increased from 0.979 to 0.984. The improvement can be credited to the incorporation of the CA module and transformer encoder block. By integrating the small target prediction head, the accuracy witnessed a significant enhancement, achieving an amazing 0.991. The inclusion of the EIoU loss function resulted in a marginal enhancement in accuracy. TCLE–YOLO has a remarkable overall average accuracy of 0.992, beating YOLOv5 by a significant margin of 1.74%. In addition, TCLE-YOLO exhibited improved performance in terms of mAP@0.5:0.95, precision (P), and recall (R) values when compared to the original YOLOv5.

TABLE I
Using the self-built dataset, we compare the evaluation indices of yolov5s with various modules.

Model	**P (%)**	**R (%)**	**mAP@0.5 (%)**	**mAP@0.5:0.9 (%)**
YOLOv5	97.90	98.10	97.50	64.30
YOLOv5 + transformer + CA module + 4 prediction heads + EIoU (TCLE–YOLO)	99.20	99.10	99.20	72.20
YOLOv5 + transformer + CA module	98.40	98.40	98.50	67.20
YOLOv5 + transformer	98.20	98.30	98.20	64.50
YOLOv5 + transformer + CA module + 4 prediction heads	99.10	99.00	99.00	71.20

Figures 6, 7, and 8 display portions of the detection and counting outcomes generated by YOLOv5s using various modules, employing the self-constructed dataset. In order to enhance clarity, the image sections that were encompassed in red boxes were enlarged, and arrows were used to identify the highlighted graphs. The grain target is denoted by the centroid of a predicted bounding box that matches to the target in the image. The number of central points is a dependable indicator of the quantity of rice grains captured in the photograph. The TCLE-YOLO model, as described in this paper, exhibited enhanced robustness and precision in rice grain counting, even under conditions of significant agglomeration. To gain a more thorough comprehension of the model's ability to detect and count, we have organised the findings in a tabular fashion, as shown in Figure 9. After examining Figure 9, it is clear that the modules of YOLOv5 exhibit a high level of competence in detecting almost all grains that are only slightly attached. Conversely, the TCLE-YOLO model distinguishes itself with its remarkable capacity to precisely quantify heavily attached rice grains. The incorporation of multiple elements, such as a transformer, a CA module, four prediction heads, and EIoU loss, led to the attainment of optimal performance. This emphasizes the efficacy of integrating these diverse modules, which are essential for permitting accurate measurements of thousand-grain weight.

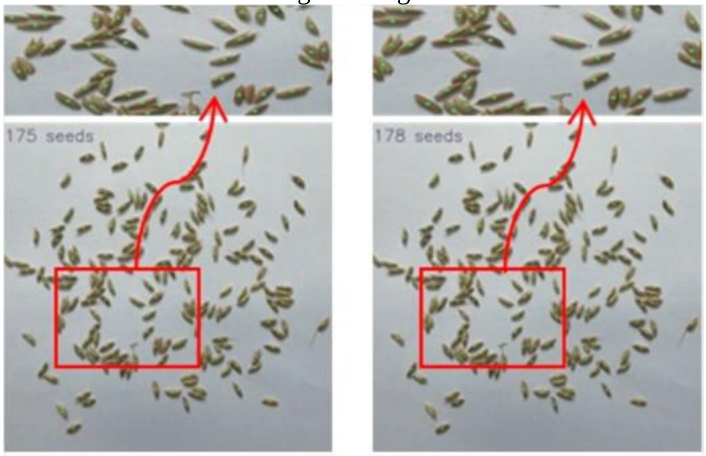


Figure 6. YOLOV5 counting and detecting result.

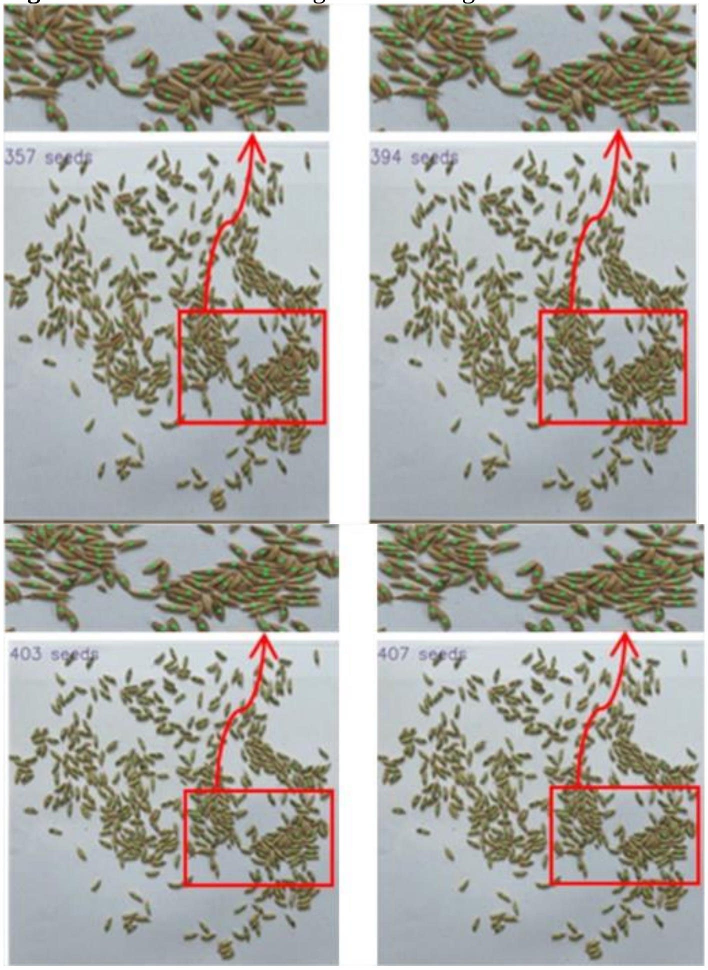


Figure 7. YOLOv5 + transformer detecting and counting results.

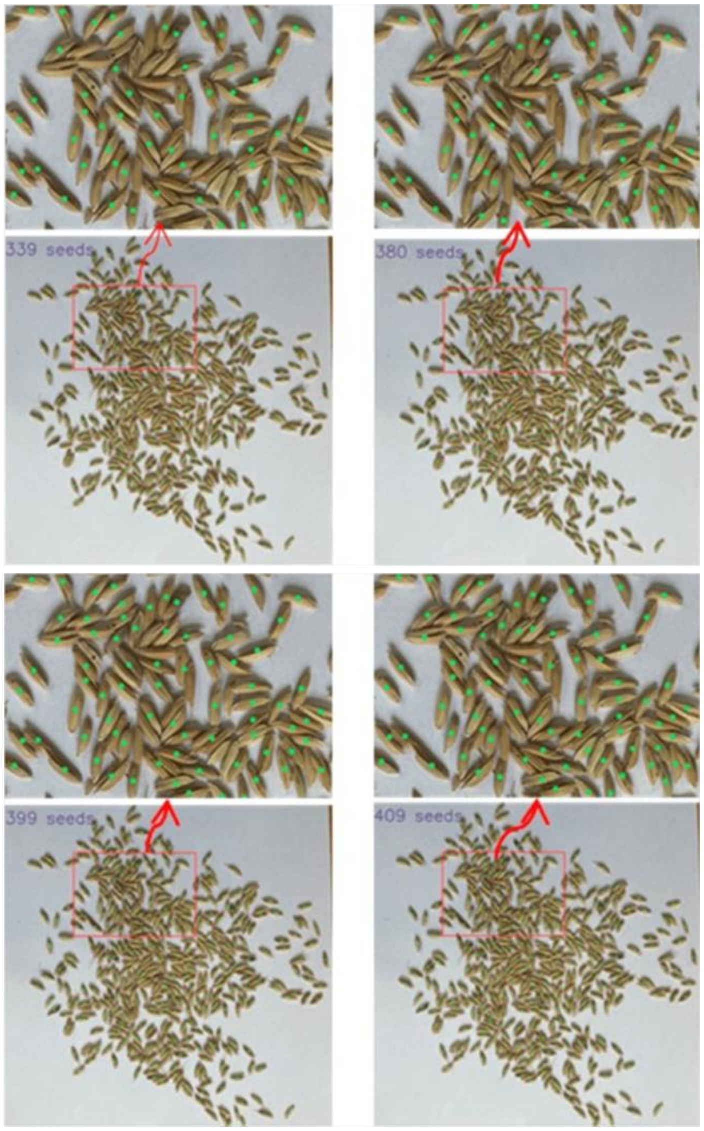
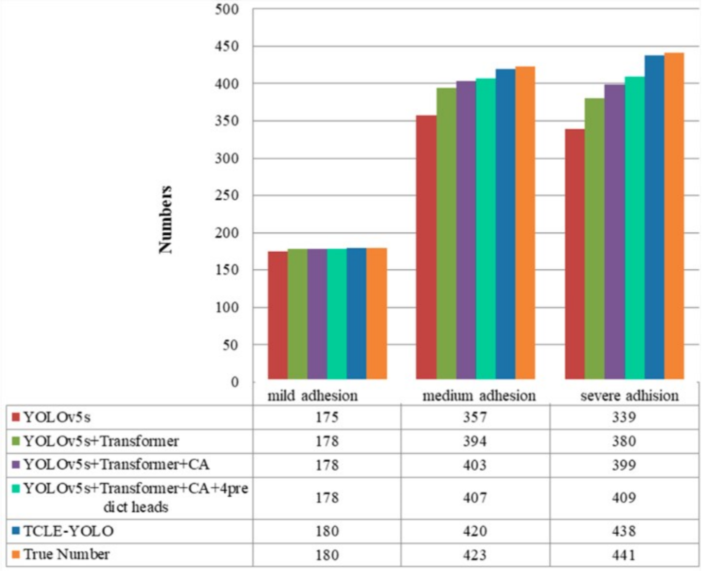


Figure 8. Assessing the efficacy of several YOLOv5s modules in identifying and counting highly sticky rice grains



	mild adhesion	medium adhesion	severe adhision
🟥 YOLOv5s	175	357	339
🟩 YOLOv5s+Transformer	178	394	380
🟪 YOLOv5s+Transformer+CA	178	403	399
🟩 YOLOv5s+Transformer+CA+4predict heads	178	407	409
🟦 TCLE-YOLO	180	420	438
🟧 True Number	180	423	441

Figure 9. Result of YOLOv5 with different modules

B. Various detection model comparison

In order to assess the accuracy of the proposed model in detecting objects, a comprehensive comparison was conducted with four other widely recognised detection models: Faster R-CNN [29], SSD [30], EfficientDet [31], and YOLOv7 [32]. The experiment utilised the same dataset, loss function, and evaluation measures as previously defined in order to maintain consistency. Table 2 presents the detection capabilities of the different models. The findings presented in Table 2 showcase the remarkable efficacy of the proposed model in comparison to the other four models in the domain of rice grain detection. The suggested model exhibited superior performance compared to Faster R-CNN in terms of precision, recall, mAP@0.5, and mAP@0.5:0.9. EfficientDet exhibited marginally higher precision and mAP@0.5:0.9, but its other two evaluation metrics were somewhat worse. SSD exhibited superior performance in comparison to Faster R-CNN and EfficientDet, albeit with somewhat lower mAP@0.5 and mAP@0.5:0.9 values. The performance of YOLOv7 was commendable, however it fell short compared to the proposed model, particularly in terms of mAP@0.5:0.9.

The visual comparisons depicted in Figure 10,11,12 and 13 offer a distinct illustration of the detection outcomes attained by the five algorithms. The graphics in the first, second, and third columns depict the ultimate predicted outcomes for grains with varying degrees of adhesion in each model. The TCLE-YOLO model exhibited its competence in precisely recognising and enumerating rice grains with varying levels of adhesion. This demonstrates the ability of the YOLOv5 backbone module to gather information on rice grains with different levels of stickiness, thanks to the inclusion of the channel attention mechanism. The suggested model's detection head prioritises the utilisation of a high-resolution feature map created by the transformer encoder at a low level. This method has significantly enhanced the process of identifying important characteristics, leading to a greater ability to detect grains that are very sticky. However, it was observed that the Faster R-CNN, EfficientDet, SSD, and YOLOv7 models were not sufficiently adaptable to extremely sticky grains. These models exhibited erroneous identifications and failed to detect certain instances, as evidenced by Figure 13c. The TCLE–YOLO model exhibited greater performance in comparison to other models, delivering a more accurate prediction of the actual quantity of rice grains.

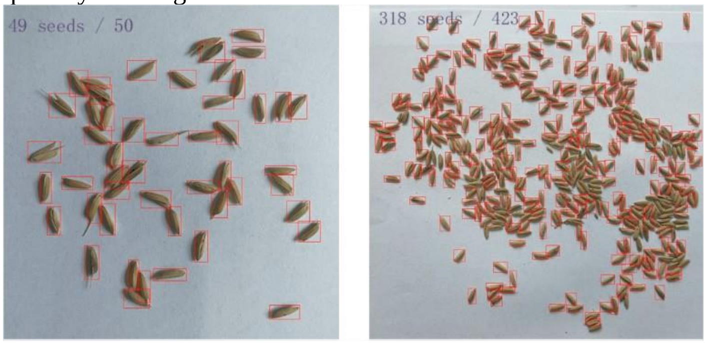


Figure 10. Faster RCNN test Result

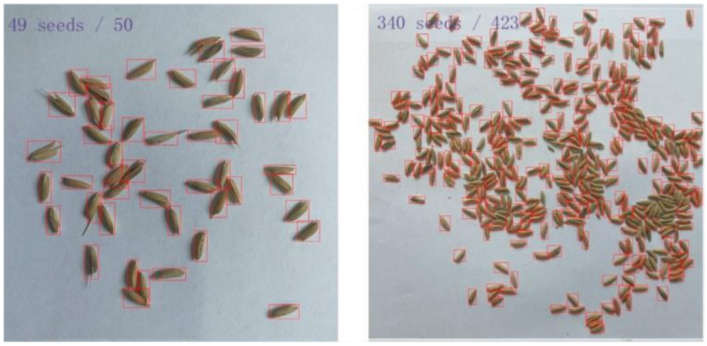


Figure 11. Efficient Net test Result

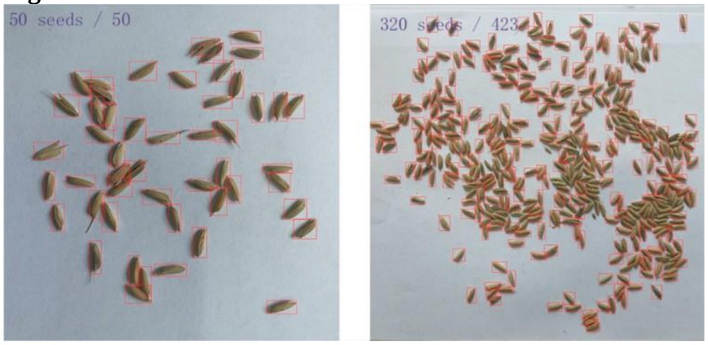


Figure 12. SSD test Result

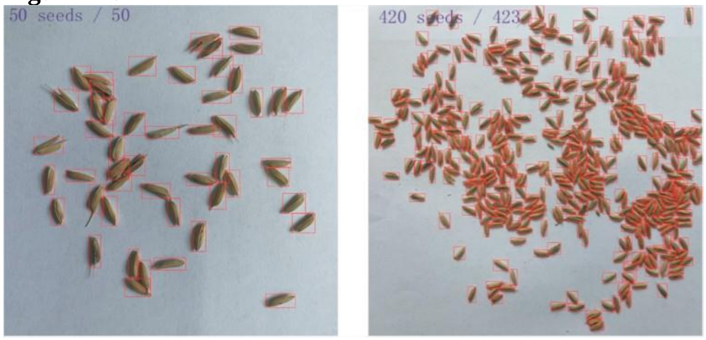


Figure 13. TCLE-YOLO test Result

IV. Discussion

Precise determination of the thousand-grain weight of rice is crucial for estimating yields, assessing cultivation techniques, and offering vital assistance for rice research. Precisely identifying and tallying rice grains can be a difficult task because of their diminutive size, limited clarity, and inconsistent levels of stickiness. We present a refined version of the YOLOv5 model that integrates an attention mechanism and an extra detecting head specifically designed for small targets. Our experiments yielded enhanced accuracy and recall metrics. Despite generating generally good findings, there were several cases where a few detections and counts were missed due to considerable occlusion between rice grains.

To address the issues we have in our prospective projects, our goal is to create an advanced picture acquisition platform that integrates two top-notch cameras and an automatic motion system. The purpose of this platform is to simplify the process of acquiring accurate images and resolving occlusion issues by utilising binocular cameras and multi-view imaging techniques. In addition, our research entailed utilising a dataset that we curated internally. In the future, our main focus will be on enhancing the model's capacity to be applicable to a broader spectrum of situations. We want to create a hardware platform specifically designed for the detection and counting of grains. Our objective is to smoothly integrate an improved model into the device, allowing for its practical use in real-world scenarios.

The model's advanced technology allows for its use to a wide range of small-seed cereals, beyond the limitation to only rice grains. This presents opportunities for detection and quantification in various other domains. This technology has the potential to completely transform agriculture processes by providing precise and automated grain detection. This tool can aid growers and breeders in estimating crop yields, assessing cultivation methods, and provide dependable data assistance for measuring phenotypes.

V. CONCLUSION

The TCLE-YOLO detection model, presented in our research, is built around an enhanced YOLOv5 framework. This model has been specifically engineered to precisely identify rice grains. In order to address problems associated with the incorrect detection of rice grains, particularly in cases where sticking together is a significant issue, YOLOv5 was improved by integrating an attention mechanism. In addition, a novel detection apparatus was specifically designed to identify and locate diminutive targets. The model underwent training, validation, and testing stages using a dataset specifically tailored for this purpose. The final test set yielded outstanding results, demonstrating great performance in terms of precision, recall, and mean average precision scores. The precision score attained a remarkable 99.20%, whereas the recall score reached 99.10%. With an IoU threshold of 0.5, the mean average precision achieved an exceptional performance of 99.20%. In addition, the model exhibited an impressive mean average precision of 72.20% over IoU thresholds ranging from 0.5 to 0.9.

Furthermore, TCLE-YOLO exhibited outstanding performance in precisely recognising and quantifying rice grains with varying degrees of adhesion, surpassing other models such as Faster R-CNN, EfficientDet, SSD, and YOLOv7. These studies showcase the effectiveness of the suggested approach in precisely identifying and measuring rice grains under different circumstances. The potential of this model for practical applications is vast, including the measurement of thousand-grain weight, the enhancement of rice breeding programmes, and the improvement of cultivation management. Future efforts will focus on enhancing the model's capacity to be applicable across diverse circumstances and extending its use to detect and measure different categories of small-seed grains.

REFERENCES

[1] Bascon, M.; Nakata, T.; Shibata, S.; Takata, I.; Kobayashi, N.; Kato, Y.; Inoue, S.; Doi, K.; Murase, J.; Nishiuchi, S. Estimating yield-related traits using UAV-derived multispectral images to improve rice grain yield prediction. Agriculture 2022, 12, 1141. [Google Scholar] [CrossRef]

[2] Liu, T.; Chen, W.; Wang, Y.; Wu, W.; Sun, C.; Ding, J.; Guo, W. Rice and wheat grain counting method and software development based on Android system. Comput. Electron. Agric. 2017, 141, 302–309. [Google Scholar] [CrossRef]

[3] Wu, W.; Zhou, L.; Chen, J.; Qiu, Z.; He, Y. GainTKW: A measurement system of thousand kernel weight based on the android platform. Agronomy 2018, 8, 178. [Google Scholar] [CrossRef]

[4] Tan, S.; Ma, X.; Mai, Z.; Qi, L.; Wang, Y. Segmentation and counting algorithm for touching hybrid rice grains. Comput. Electron. Agric. 2019, 162, 493–504. [Google Scholar] [CrossRef]

[5] Liu, S.; Yin, D.; Feng, H.; Li, Z.; Xu, X.; Shi, L.; Jin, X. Estimating maize seedling number with UAV RGB images and advanced image processing methods. Precis. Agric. 2022, 23, 1604–1632. [Google Scholar] [CrossRef]

[6] Wu, W.; Liu, T.; Zhou, P.; Yang, T.; Li, C.; Zhong, X.; Guo, W. Image analysis-based recognition and quantification of grain number per panicle in rice. Plant Methods 2019, 15, 1–14. [Google Scholar] [CrossRef]

[7] Kumar, J.P.; Domnic, S. Image based leaf segmentation and counting in rosette plants. Inf. Process. Agric. 2019, 6, 233–246. [Google Scholar]

[8] Wu, Z.; Sun, X.; Jiang, H.; Mao, W.; Li, R.; Andriyanov, N.; Soloviev, V.; Fu, L. NDMFCS: An automatic fruit counting system in modern apple orchard using abatement of abnormal fruit detection. Comput. Electron. Agric. 2023, 211, 108036. [Google Scholar] [CrossRef]

[9] Sarijaloo, F.B.; Porta, M.; Taslimi, B.; Pardalos, P.M. Yield performance estimation of corn hybrids using machine learning algorithms. Artif. Intell. Agric. 2021, 5, 82–89. [Google Scholar] [CrossRef]

[10] Fuentes-Peñailillo, F.; Carrasco Silva, G.; Pérez Guzmán, R.; Burgos, I.; Ewertz, F. Automating seedling counts in horticulture using computer vision and AI. Horticulturae 2023, 9, 1134. [Google Scholar] [CrossRef]

[11] Khaki, S.; Pham, H.; Han, Y.; Kuhl, A.; Kent, W.; Wang, L. Convolutional neural networks for image-based corn kernel detection and counting. Sensors 2020, 20, 2721. [Google Scholar] [CrossRef] [PubMed]

[12] Gong, L.; Fan, S. A CNN-Based Method for Counting Grains within a Panicle. Machines 2022, 10, 30. [Google Scholar] [CrossRef]

[13] Tan, C.; Li, C.; He, D.; Song, H. Towards real-time tracking and counting of seedlings with a one-stage detector and optical flow. Comput. Electron. Agric. 2022, 193, 106683. [Google Scholar] [CrossRef]

[14] Lyu, S.; Li, R.; Zhao, Y.; Li, Z.; Fan, R.; Liu, S. Green citrus detection and counting in orchards based on YOLOv5-CS and AI edge system. Sensors 2022, 22, 576. [Google Scholar] [CrossRef] [PubMed]

[15] Rong, J.; Zhou, H.; Zhang, F.; Yuan, T.; Wang, P. Tomato cluster detection and counting using improved YOLOv5 based on RGB-D fusion. Comput. Electron. Agric. 2023, 207, 107741. [Google Scholar] [CrossRef]

[16] Koklu, M.; Cinar, I.; Taspinar, Y.S. Classification of rice varieties with deep learning methods. Comput. Electron. Agric. 2021, 187, 106285. [Google Scholar] [CrossRef]

[17] Yang, B.; Gao, Z.; Gao, Y.; Zhu, Y. Rapid detection and counting of wheat ears in the field using YOLOv4 with attention module. Agronomy 2021, 11, 1202. [Google Scholar] [CrossRef]

[18] Sun, P.; Chen, G.; Cao, L. Image recognition of soybean pests based on attention convolutional neural network. J. Chin. Agric. Mech. 2020, 41, 171–176. [Google Scholar]

[19] Zhang, W.; Ma, H.; Li, X.; Liu, X.; Jiao, J.; Zhang, P.; Gu, L.; Wang, Q.; Bao, W.; Cao, S. Imperfect wheat grain recognition combined with an attention mechanism and residual network. Appl. Sci. 2021, 11, 5139. [Google Scholar] [CrossRef]

[20] Wen, G.; Li, S.; Liu, F.; Luo, X.; Er, M.J.; Mahmud, M.; Wu, T. YOLOv5s-CA: A Modified YOLOv5s Network with Coordinate Attention for Underwater Target Detection. Sensors 2023, 23, 3367. [Google Scholar] [CrossRef]

[21] Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13713–13722. [Google Scholar]

[22] Tian, Y.L.; Wang, Y.T.; Wang, J.G.; Wang, X.; Wang, F.Y. Key problems and progress of vision Transformers: The state of the art and prospects. Acta Autom. Sin. 2022, 48, 957–979. [Google Scholar]

[23] Patrício, D.I.; Rieder, R. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. Comput. Electron. Agric. 2018, 153, 69–81. [Google Scholar] [CrossRef]

[24] Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. J. Big Data 2019, 6, 1–48. [Google Scholar] [CrossRef]

[25] Wang, X.; Yang, W.; Lv, Q.; Huang, C.; Liang, X.; Chen, G.; Duan, L. Field rice panicle detection and counting based on deep learning. Front. Plant Sci. 2022, 13, 966495. [Google Scholar] [CrossRef] [PubMed]

[26] Lang, X.; Ren, Z.; Wan, D.; Zhang, Y.; Shu, S. MR-YOLO: An improved YOLOv5 network for detecting magnetic ring surface defects. Sensors 2022, 22, 9897. [Google Scholar] [CrossRef]

[27] Wu, X.; Doyen, S.; Steven, C.H. Recent advances in deep learning for object detection. Neurocomputing 2020, 396, 39–64. [Google Scholar] [CrossRef]

[28] Li, W.; Liu, C.; Chen, M.; Zhu, D.; Chen, X.; Liao, J. A lightweight semantic segmentation model of Wucai seedlings based on attention mechanism. Photonics 2022, 9, 393. [Google Scholar] [CrossRef]

[29] Wu, W.; Yang, T.L.; Li, R.; Chen, C.; Liu, T.; Zhou, K.; Sun, C.M.; Li, C.Y.; Zhu, X.K.; Guo, W.S. Detection and enumeration of wheat grains based on a deep learning method under various scenarios and scales. J. Integr. Agric. 2020, 19, 1998–2008. [Google Scholar] [CrossRef]

[30] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.; Berg, A. Ssd: Single shot multibox detector. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37. [Google Scholar]

[31] Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790. [Google Scholar]

[32] Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. 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, Oxford, UK, 15–17 September 2023; pp. 7464–7475.

AUTHORS

First Author – Faiq Shah is currently pursuing his Master Degree in Mechatronics Engineering From UET Peshawar
Second Author – Kamran Shah is Currently working as Associate Professor at Department of Mechatronics Engineering UET Peshawar.
Third Author – Izhar Ul Haq is Currently working as Professor at Department of Mechatronics Engineering UET Peshawar Moreover he is the Principal Investigator of Advance robotics and Automation Lab.

Correspondence Author – **Faiq Shah**