PaddyFormer: An Improved RT-DETRv2 based Approach for Paddy Crop Growth Stage Detection on Drone based RGB Imagery

Anonymous submission

Abstract

Accurately recognizing crop growth stages is vital in precision agriculture, particularly for predicting yield and determining harvesting times. However, this task is challenging due to the significant morphological variations across different growth stages, often impacting model performance. This study addresses the five-class growth stage recognition task for paddy crops, using high-resolution drone-based RGB imagery captured by a DJI Inspire-1 Pro drone equipped with a Zenmuse X5 camera. We propose PaddyFormer, an enhanced version of RT-DETRv2, integrated with a weighted dataloader and asymmetric loss to handle class imbalance and field-level variability effectively. Our experimental results show that the proposed approach, PaddyFormer, demonstrates strong performance, achieving the highest $mAP@[0.5]$ of 84.5%, with a precision of 75.6% and a recall of 82.5%, highlighting its effectiveness and robustness under complex agricultural conditions. Overall, this emphasizes the importance of drone-based image acquisition, transformer-based in developing scalable, real-time solutions for crop monitoring.

Introduction

More than 50% of the world's population rely on paddy (*Oryza sativa L.*) as a staple food. With a global production of 756 million tonnes, it is the third most widely cultivated crop (Food and Organisation 2019). Accurate information about the growth stages, such as the timing of flowering and harvest, is essential to implement timely, tailored strategies for effective crop management. Crops undergo continuous morphological changes throughout their growth cycle, enabling the use of advanced technologies to automatically detect various growth stages. Developing robust plant recognition systems that handle this morphological variation is essential to advance high-throughput plant phenotyping and precision agriculture technologies (Danilevicz et al. 2021). Computer vision, integrated with remote sensing technologies such as drones/UAVs (Unmanned Aerial Vehicles), enables large-scale, high-resolution crop monitoring across seasons.

Deep Learning (DL) has revolutionized machine learning by eliminating the need for manual feature engineering and enabling automatic feature extraction. DL-based object detection models are typically categorized into one-stage and two-stage approaches. Two-stage models, such as Faster R-CNN (Ren et al. 2015), often achieve high accuracy but suffer from slower inference speeds due to sequential processing. In contrast, one-stage models like YOLO (You Only Look Once) (Redmon 2016) offer a better balance between accuracy and speed, although they encounter challenges with Non-Maximum Suppression (NMS). While DETR models address the NMS limitations inherent with YOLO, their high computational demands hinder their ability to perform real-time detection, offsetting the advantages of an NMS-free design (Carion et al. 2020). To bridge this gap, the authors in (Zhao et al. 2024) introduced RT-DETR (Real-Time DEtection TRansformer). RT-DETR is built on DETR with an efficient hybrid encoder and allows flexible speed-accuracy trade-offs without retraining. RT-DETRv2 (Lv et al. 2024) is further improved with improvements such as scale-specific sampling in deformable attention for enhanced multiscale feature extraction, replacing grid sampling with discrete operators for easier deployment, and employing dynamic augmentation along with adjustable hyperparameters to boost efficiency while maintaining real-time capabilities.

Various studies have developed DL-based techniques to monitor the growth stages of paddy. The authors in (Tan et al. 2022) found that EfficientnetB4 achieved the best performance on UAV-captured data to detect rice seedling growth stages. Zhou et al. (Zhou et al. 2023) designed a multitask pipeline of YOLOv5, ResNet50, and DeepSORT to detect and track flowering panicles collected by a static pole-mounted camera system and found that YOLOv5 is more robust to background noise than the R-CNN models. Chen et al. (Chen et al. 2023) used UAV imagery to evaluate object detection models to count rice panicles, identifying YOLOv8-X as the most effective for monitoring early and late heading stages.

Despite significant advancements, existing DL-based models face notable challenges regarding agricultural tasks. These models often struggle with complex visual scenarios, such as occlusion, background clutter, dense crop scenes, varying lighting conditions, and detecting small objects like weeds in the field. While prior studies have contributed to the detection of specific paddy growth stages such as seedling, flowering, or 50% heading, they essentially fall short in addressing the full spectrum of growth stages in paddy crop. Furthermore, as the crop progresses from veg-

etative to ripening, class imbalance becomes increasingly prominent. To address the issues mentioned above, this study presents PaddyFormer, an enhanced version of RT-DETRv2 for accurate paddy growth detection across multiple stages. The proposed model effectively tackled various challenges encountered in real-field scenarios. The contributions of this study are as follows:

- We curate a new drone-captured image dataset for paddy growth stage recognition, comprising 798 images of paddy crop, with 6,445 manually annotated plots. The dataset spans the entire paddy growth cycle from the vegetative to the ripening phase. All images were captured under natural lighting in real-field conditions. It will be made publicly available.
- We employ a weighted dataloader and various data augmentations to tackle the class imbalance and visual variations present in the dataset.
- We integrate asymmetric loss into the RT-DETRv2 framework to improve model learning by modulating the gradients of positive and negative samples, thereby enhancing its robustness in complex and imbalanced scenarios.

Materials and Methods

Field Preparation

The experimental study was carried out during the Kharif season in 2019, from July to November, in a semi-arid region of Hyderabad, Telangana, India. The agricultural site was managed by the Institute of Biotechnology of Professor Jayashankar Telangana State Agriculture University, located in Hyderabad, India. The study area spans 15.3 m x 34.8 m and includes two repetitions of 203 plots, each corresponding to a unique variety of aerobic paddy, resulting in 406 aerobic paddy plots. Each plot covers an area of 1.26 m^2 and contains 42 crop strands.

Image Data Collection

DJI Inspire-1 Pro drone coupled with the Zenmuse X5 camera was used to collect data. This camera features a 16-megapixel CMOS sensor with an ISO range of 100 to 25,600. To ensure high-quality and consistent imagery for downstream detection, flight trajectories were optimized using Mission Planner v4.3.1 (ArduPilot Dev Team), with missions conducted at a height of 7 meters and a speed of 4 km/h. These settings, combined with sensor calibration and environmental controls, resulted in an effective ground sampling distance of 3.2 to 4 centimeters. To ensure comprehensive coverage, consecutive images were captured with a horizontal overlap of 50-70% and a vertical overlap of 70-80%. Weekly data acquisition was conducted in consultation with agricultural scientists to align with distinct paddy crop growth phases, whose durations are summarized in Table 1. The weed class was also included to improve early-stage discrimination, especially against seedlings. The final dataset consists of six classes—five crop growth stages and one weed class as visualized in Figure 1.

Image Annotation

To minimize redundancy, an image was selected from every four to five successive UAV-captured raw images, resulting in a total of 793 images. These images were annotated using the open-source tool Label Studio (Tkachenko et al. 2020-2022) in the standard YOLO format, producing 6,445 annotated instances of paddy plots across all growth stages. Each image was examined and corrected by an expert annotator under the guidance of a professional specialized in genetics and plant breeding. The number of instances for each category is presented in Table 1.

Figure 1: Example sample from each class of the dataset - Weed, Seedling, Tillering, Booting, Flowering, and Harvesting.

PaddyFormer: Our Proposed Method

Our proposed method, PaddyFormer, is an enhanced version of RT-DETRv2 (Lv et al. 2024), designed to address the class imbalance and improve detection performance in complex real-field scenarios. While RT-DETRv2 itself advances over RT-DETR (Zhao et al. 2024) by introducing a query instantiator for instance-level feature interaction, reducing decoder depth without sacrificing accuracy, and improving training stability for real-time applications, it does not explicitly tackle class imbalance issues. To overcome this, PaddyFormer, shown in Figure 2, integrates a weighted dataloader to ensure balanced learning across underrepresented classes. Additionally, we incorporate an asymmetric loss function into the RT-DETRv2 framework, further strengthening the model's ability to focus on hard and minority samples, thereby enhancing robustness and accuracy in challenging field conditions.

Weighted Dataloader Technique The dataset used in this study shows a class imbalance, with certain classes being underrepresented. Various strategies have been explored to address the class imbalance in object detection tasks, including sampling methods, loss weighting, and data augmentation (Crasto 2024). Unlike techniques that rely on undersampling majority classes or applying class-specific loss weights, we adopt a weighted dataloader approach that

Table 1: Overview of growth phases and growth stages in paddy.

Growth Phase	Growth Stage	Time Period (days)	Description	Class ID	Instances	Bounding Box Area ($pixel^2$)
Weed	Weed	-	Unwanted plants in the field	Weed	1180	7.35±21.61
Vegetative	Seedling	10-15	Leaf node with <6 leaves per each sapling	VE-st1	1367	114.28±85.25
	Tillering	45-75	Multiple leaf nodes, lush green canopy	VE-st2	1049	146.832±69.51
Reproductive	Booting	90	Beginning of flag leaves	RE-st1	967	117.28±57.54
	Flowering	90-120	Presence of panicles throughout the section	RE-st2	823	157.87±85.76
Ripening	Harvesting	120-150	Fully ripened heavy panicles leaning towards the ground	RI-st1	1059	186.64±106.63

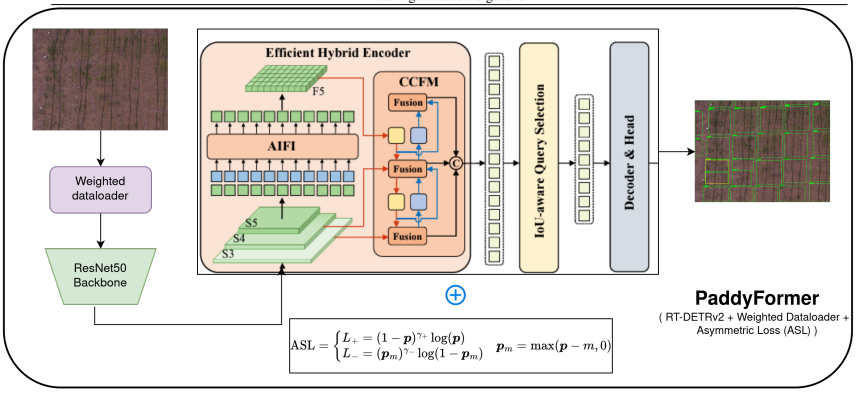

Figure 2: Overview of Proposed Method - PaddyFormer, built of RT-DETRv2 (Lv et al. 2024) integrated with a weighted dataloader and asymmetric loss(Ridnik et al. 2021).

maintains the integrity of the dataset while promoting balanced learning. This method computes sampling probabilities based on the inverse frequency of each class, ensuring that images containing underrepresented classes are selected more frequently during training. A reduction of 45% in the variance of class representation was observed after incorporating the weighted dataloader technique into the training pipeline, indicating a significant improvement in class balance (shown in Figure 3). Notably, the weighted dataloader achieves this without discarding any samples or introducing overfitting risks, thereby preserving data diversity and enhancing the robustness of the trained model.

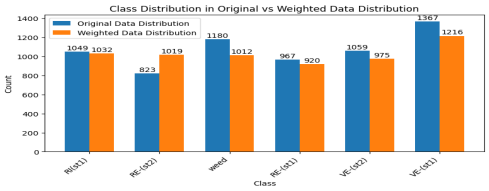

Figure 3: Class Distribution in Original vs Weighted distributions on Training data

Asymmetric Loss Function In paddy growth stage detection, predictions can be both accurate and inaccurate. Positive samples refer to correctly detected instances, while negative samples represent incorrect detections. In complex field environments, the model tends to generate more negative and complex samples, which can overwhelm the learning process and hinder the model's ability to focus on positive samples, thereby reducing overall detection accuracy. To mitigate this, we propose using Asymmetric Loss, introduced by Ridnik et al. (Ridnik et al. 2021), for imbalanced multi-label classification tasks, instead of traditional cross-entropy loss given in equation **??** for the confidence branch. The Asymmetric Loss function in equation 2 applies different weights to positive and negative samples to better balance their influence. It calculates the loss using the logarithmic function for both positive and negative predictions based on the predicted probability. The loss formulation includes two focusing parameters, γ_+ and γ_- control the weighting of positive and negative samples, respectively. Additionally, a hard threshold, m (also referred to as the clip value), is used to suppress easy negatives by clipping the predicted probabilities of negative samples to a minimum value of m. In our implementation, the Asymmetric Loss function is defined with $\gamma_- = 4$ and $\gamma_+ = 1$, which allows the model to focus more on the negative samples and less on the easy positive samples and m, is set to 0.05, which helps to suppress easy negatives.

$$\text{CE} = \begin{cases} L_+ = \log(p) \\ , L_- = \log(1-p), \end{cases} \tag{1}$$

$$\text{ASL} = \begin{cases} L_+ = (1-\boldsymbol{p})^{\gamma_+}\log(\boldsymbol{p}), \\ L_- = (\boldsymbol{p}_m)^{\gamma_-}\log(1-\boldsymbol{p}_m) \end{cases} \qquad \boldsymbol{p}_m = \max(\boldsymbol{p}-m, 0) \tag{2}$$

Model Training

The dataset for each growth stage was divided into training and validation sets in an 8:2 ratio. We applied various data augmentation techniques to training data to account for inherent variations in the dataset, such as differences in texture, illumination, and visual characteristics across growth

stages. These included random brightness, random contrast, CLAHE (Contrast Limited Adaptive Histogram Equalization), gaussian blur, mosaic, and shear transformations. These augmentations enhanced the models' generalizability by providing diverse data perspectives without altering the ground truth counts. To address class imbalance, we employed a weighted dataloader technique, as described in section . We fine-tuned the models using the default hyperparameters specified in the RT-DETRv2 repository[1]. The training process was stopped when validation loss consistently increased, indicating the onset of overfitting. The training process was conducted on a system equipped with a 32-core Intel(R) Xeon(R) Silver 4110 CPU and an NVIDIA Tesla V100 SXM3 GPU with 32 GB of RAM, running Ubuntu 20.04 and PyTorch framework. All data, model configurations, and best weights for all trained variants will be made available.

Evaluation Metrics

In this study, the performance of object detection models was evaluated using three metrics such as precision, recall, and mAP@[0.5](mean Average Precision at threshold of 0.5) defined by equations 3, 4 and 5. Precision, which assesses the model's ability to predict correct positive outcomes, while recall assesses the model's ability to capture all positive cases in the dataset. FN denotes False Negatives, TP denotes True Positives, and FP indicates False Positives with respect to the actual and predicted classes.

$$Precision = \frac{TP}{(TP + FP)} \tag{3}$$

$$Sensitivity/Recall = \frac{TP}{(TP + FN)} \tag{4}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} \text{AP}_i \tag{5}$$

Results and Discussions

Model Performance Results

The results of the proposed method PaddyFormer are compared with various state-of-the-art object detection models and presented in Table 2. PaddyFormer achieves the best overall performance, with a precision of 0.756, recall of 0.825, and $mAP@[0.5]$ of 84.5%. This superior performance can be attributed to its RT-DETRv2-based architecture, which is customized to address class imbalance present in the dataset and enhance the model learning by integrating asymmetric loss. Notably, PaddyFormer outperforms the vanilla RT-DETRv2 and RT-DETR models, demonstrating the effectiveness of its design. Although PaddyFormer has the largest parameter size of 76 million, its performance justifies the complexity. In contrast, YOLO variants such as YOLOv10-X and YOLOv8-L achieve competitive performance, with an $mAP@[0.5]$ of around 82%, using fewer parameters when compared with RT-DETRv2 variants. Two-stage detectors, including Faster RCNN and RetinaNet, struggle to capture the dataset's complexities, resulting in poor performance.

[1](https://github.com/lyuwenyu/RT-DETR/tree/main/rtdetrv2_pytorch)

Understanding the relationship between growth stage and model performance

The performance of the proposed method and various state-of-the-art object detection models across all the growth stages are reported in Table 3. It can be observed that the proposed method PaddyFormer consistently achieves notable performance on seedling with 89.8%, tillering with 92.3%, and booting with 91.4%, outperforming other state-of-the-art models. While YOLOv10-X and YOLOv5-X show competitive results in some stages, such as booting with 91.1% and tillering with 92.6%, they fall short in small object weed detection, whereas PaddyFormer excels in the weed category with a score of 69.54%, which is the highest among all models. Traditional detectors like Faster R-CNN and RetinaNet lag significantly, particularly in later stages and under challenging conditions. The qualitative results of the proposed method in various scenes are presented in Figure 4.

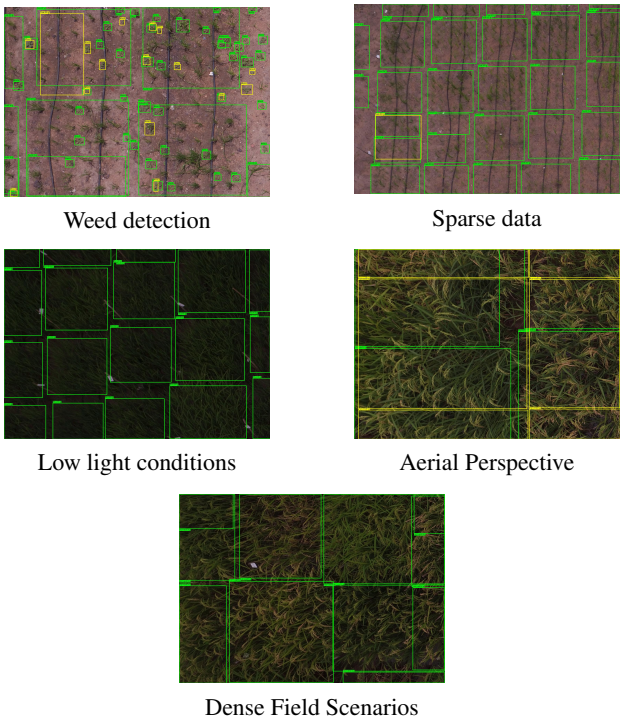

Figure 4: Illustration of growth stage detection in paddy fields using PaddyFormer under five different conditions. Green boxes represent model predictions, while yellow boxes indicate missed detections.

Conclusion

This study presents PaddyFormer, an enhanced version of RT-DETRv2, designed to tackle the challenges of class imbalance and complex scenes in agricultural field conditions. By incorporating a weighted dataloader and asymmetric

Table 2: Comparison of growth stage recognition performance with different models, Results **bolded** indicate the overall top performing model.

Version	*Precision (↑)*	*Recall (↑)*	*mAP@0.5 (↑)*	*mAP@[0.5:0.95] (↑)*	*Time [hrs] (↓)*	*GFLOPS (↓)*	*Params (million) (↓)*
Faster-RCNN (Ren et al. 2015)	0.63	0.79	63.01	33.41	0.5	85.58	41.3
RetinaNet(Lin et al. 2017)	0.75	0.48	41.03	23.65	0.5	60.18	36.32
YOLOv5-X(Jocher 2020)	0.737	0.794	81.7	47.2	1.007	246	97.15
YOLOv8-L(Jocher, Chaurasia, and Qiu 2023)	0.725	0.808	82.2	47.2	0.625	164.8	43.61
YOLOv10-X(Wang et al. 2024)	0.739	0.811	82.3	48	1.029	206.1	31.59
RT-DETR-L(Zhao et al. 2024)	0.737	0.77	76.3	44.6	0.985	103.5	31.99
RT-DETRv2-X(Lv et al. 2024)	0.734	0.74	77.57	50.14	2.4	25.9	76
PaddyFormer (Ours)	**0.756**	**0.825**	**84.5**	**51.12**	3.21	26.1	76

Table 3: Performance of models on individual growth stage classes in terms of *mAP@0.5*. Bold values indicate the best performing model of that growth stage.

Growth stage	**Faster-RCNN**	**RetinaNet**	**YOLOv5-X**	**YOLOv8-L**	**YOLOv10-X**	**RT-DETR-L**	**RT-DETRv2-X**	**PaddyFormer (Ours)**
Seedling	55.6	60.2	89.5	87.7	89.6	88.2	87.9	**89.8**
Tillering	72.1	71.5	92.3	90.4	**92.6**	90.5	79.9	92.3
Booting	63.5	54.9	87.8	85.8	91.1	87.3	80.5	**91.4**
Flowering	80.6	79.4	83.9	**87.6**	84.6	86.3	69.9	81.92
Harvesting	73.5	74.4	70.8	75.6	73.7	76.27	72.2	**77.9**
Weed	27.7	28.9	66.1	66.3	29.7	61.9	65.4	**69.54**

loss, PaddyFormer achieves robust performance in detecting paddy growth stages. Our comprehensive evaluation demonstrates PaddyFormer's superiority, with an $mAP@[0.5]$ of 84.5%, precision of 75.6%, and recall of 82.5%. Although computationally intensive compared to recent YOLO variants and RT-DETR, the performance of the proposed method justifies its complexity. To our knowledge, this is the first study to apply such an approach in the Indian agricultural context, offering valuable insights for high-throughput phenotyping and paving the way for future research in crop monitoring and management.

References

Carion, N.; et al. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, R.; Lu, H.; Feng, Q.; and Han, B. 2023. High-throughput UAV-based rice panicle detection and genetic mapping of heading-date-related traits. *Frontiers in Plant Science*, 15: 1327507.

Crasto, N. 2024. Class imbalance in object detection: an experimental diagnosis and study of mitigation strategies. *arXiv preprint arXiv:2403.07113*.

Danilevicz, M. F.; Bayer, P. E.; Nestor, B. J.; Bennamoun, M.; and Edwards, D. 2021. Resources for image-based high-throughput phenotyping in crops and data sharing challenges. *Plant physiology*, 187(2): 699–715.

Food; and Organisation, A. 2019. FAO Cereal Supply and Demand Brief Report.

Jocher, G. 2020. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements. https://github.com/ultralytics/yolov5.

Jocher, G.; Chaurasia, A.; and Qiu, J. 2023. Ultralytics YOLOv8.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lv, W.; Zhao, Y.; Chang, Q.; Huang, K.; Wang, G.; and Liu, Y. 2024. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer. *arXiv preprint arXiv:2407.17140*.

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

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; and Zelnik-Manor, L. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 82–91.

Tan, S.; Liu, J.; Lu, H.; Lan, M.; Yu, J.; Liao, G.; Wang, Y.; Li, Z.; Qi, L.; and Ma, X. 2022. Machine learning approaches for rice seedling growth stages detection. *Frontiers in Plant Science*, 13: 914771.

Tkachenko, M.; Malyuk, M.; Holmanyuk, A.; and Liubimov, N. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Wang, A.; et al. 2024. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*.

Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; and Chen, J. 2024. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16965–16974.

Zhou, Q.; Guo, W.; Chen, N.; Wang, Z.; Li, G.; Ding, Y.; Ninomiya, S.; and Mu, Y. 2023. Analyzing nitrogen effects on rice panicle development by panicle detection and time-series tracking. *Plant Phenomics*, 5: 0048.