International Journal of Agriculture and Animal Production
Vol: 05, No. 01, Jan-June 2025
ISSN: 2799-0907, DOI: 10.55529/ijaap.51.129.144

Research Paper

Robust visual navigation and adaptive control decision-making for autonomous agricultural robots in mature wheat fields: an improved RT-DETRv2 and Fuzzy-PID framework

Madhusmita Swain*
*Research Scholar, Kalinga University, Raipur & Postgraduate Teacher (Zoology), Govt. Higher Secondary School, Tudalaga, Sundargarh, Odisha., India.

Article Info

Article History:
Received: 16 January 2025
Revised: 25 March 2025
Accepted: 02 April 2025
Published: 19 May 2025

Keywords:
Visual Navigation
Wheat Crop Row Detection
RT-DETRv2
Multi-Scale Attention
Adaptive Fuzzy-PID
Precision Agriculture

ABSTRACT

Perceptual challenges are extreme in mature wheat due to dense canopy occlusion, specular reflections from senescent awns, wind lodging and therefore variable inter-row spacing, and the challenge of having to operate a combine in real time. In this situation, the authors propose a framework, RT-DETRv2-MSCA-SCGE with Adaptive Fuzzy-PID (RTMS-AFP), which combines the visual crop-row detection and real-time control decision-making for an autonomous agricultural machine. From the perceptual perspective, two modules are added: a Multi-Scale Channel Attention (MSCA) module in the transformer encoder backbone to improve the receptive field coverage and a Spatial Context Global Encoder (SCGE) module in the feature pyramid neck, to improve the edge-feature discrimination in dense canopy scenarios. The detection model was trained using a custom-made dataset of 7,846 annotated wheat-field images (2,134 images captured at original locations and 5,712 images created with diverse conditions), across seven environmental conditions in three wheat-growing regions in India and China. A PID controller with adaptive gains is implemented on the control side to adaptively adjust PID gains continuously through real-time fuzzy inference, which is derived from lateral deviation and heading angle error obtained from the detection pipeline. With a real-time processing time of 26.4 ms per frame, the proposed model achieved a mAP@50 of 97.1%, a mAP@50-95 of 89.4%, 3.84 px mean lateral error and 2.46° mean heading error, meeting real-time requirements with improvements of +2.96%, +4.93%, and -25.0% over baseline RT-DETRv2. Field trials over three wheat growing seasons in Yangling, China and Coimbatore, India resulted in an average crop row recognition accuracy of 97.8 % and RMSE of 0.041 m for forward speeds up to 1.1 m/s. One-way ANOVA ($F = 74.2$, $p < 0.001$) and Mann-Whitney U tests ($p < 0.001$) showed that the proposed framework was significantly superior to all the baseline frameworks. These findings demonstrate the ability of RTMS-AFP to solve the autonomous navigation problem in the most difficult part of the cereal harvest process, while being efficient in terms of computational power.

Corresponding Author:
Madhusmita Swain
Research Scholar, Kalinga University, Raipur & Postgraduate Teacher (Zoology), Govt. Higher Secondary School, Tudalaga, Sundargarh, Odisha., India.
Email: madhusmitaswain43@gmail.com

1. INTRODUCTION

In the last century, wheat (Triticum aestivum L.) has been the most widely cultivated cereal worldwide, grown on 218 million hectares of land [1], and is facing significant pressure from the global cereal economy due to declining availability of rural labour, increased operational costs and the agronomic need to harvest the crop in time with precision. Combine operations in dense mature wheat fields pose one of the most challenging scenarios for autonomous agricultural robots, as there are three compounding problems: (i) Extremely high inter-class appearance confusion is created by ripened awns, yellowed leaf surfaces, and specular straw surfaces of the background; (ii) The geometric inter-row spacing is variable due to wind lodging and uneven germination; and (iii) The hard real-time constraint imposed by the throughput requirements of the combine, where delayed or incorrect navigation decisions immediately lead to higher straw losses and header misalignment damage [2], [3].

With the advantages of fine-grained perceptual ability, low investment cost, and independent positioning station, the machine vision-based navigation system has become the mainstream navigation method for guiding the agricultural robot to complete autonomous navigation, replacing the former global positioning system (GPS) and ultrasonic navigation systems [4], [5]. From the region-proposal network paradigm [6] to single-stage detectors, agricultural row-detection adaptations [7] to the trans-formative attention-based DETR model [8], crop row detection models have indeed seen great evolution over the years, leading to significant increase in both accuracy and generalizability. However, there are three drawbacks in the current methods that hinder them from performing to the level necessary to achieve full autonomy of the combine in a mature wheat field.

First, there are existing crop row detection models, which are mostly evaluated in the early growth or mid-season of the crops, and background clutter is low and inter-rows contrast is high [9]. In the most critical scenario of wheat growing (maturity), visual ambiguity is present in significant ways, such as the overlap of crop canopy borders, shadows from the grain heads, and the saturated yellow-green spectral overlap that happens when the crop canopy is filled to its maximum [10].The most critical scenario of wheat growing, when the wheat is mature, exhibits severe visual ambiguity in the form of overlapping edges of crop canopy, shadows cast by the grain heads, and saturated yellow-green spectral overlap when the crop canopy is filled to its maximum. Second, transformer-based architectures like DETR [8] and RT-DETR [11] are able to better model the long-range context than their convolutional counterparts, however, their commonly used feature pyramid structures may not be strong enough to capture the edge features in fine scale, such as the crop-row boundary, that is the most important feature for navigation line localization [12] in an operationally critical scale. Third, the navigation control layer in current ag robot systems usually uses static PID controllers with a fixed operating speed that tend to suffer from large lateral overshoot and settling-time degradation when the operating conditions change, which is common in field harvesting caused by uneven terrain, changing crop density and dynamic changes in operating speed [13], [14].

In order to overcome all three of these three gaps, a RT-DETRv2-MSCA-SCGE with Adaptive Fuzzy-PID (RTMS-AFP) framework is proposed in this study. The perceptual component has been modified by introducing two new components: (i) a Multi-Scale Channel Attention (MSCA) module that fuses channel-

wise recalibration of features within three feature scales of its encoder, which are salient for the edge signatures of wheat row boundaries, which are spectrally distinct; and (ii) a Spatial Context Global Encoder (SCGE) in the feature pyramid neck that uses deformable attention to fuse long-range spatial context that is essential for resolving row discontinuities caused by lodging. The control component presents an Adaptive Fuzzy-PID controller which can tune its gains automatically in real-time based on a two-input (lateral error and heading error) fuzzy inference system to achieve tight path tracking at the entire operating speed range without manual gain adjustment.

Main contributions are: (1) it presents a systematic dataset and evaluation framework of mature wheat visual navigation, which spans across seven environmental conditions in two countries; (2) it presents two modules (MSCA, SCGE), which can be plugged-in as improvements to the RT-DETRv2 architecture and were shown to be accurate and efficient, through ablation study; (3) it presents an Adaptive Fuzzy-PID navigation controller validated by 120 independent field tests; and (4) statistical analysis (ANOVA, Mann-Whitney u, Wilcoxon signed-rank) is proposed and provides evidence of significant superiority over baseline methods, jointly optimizing both the detection module and the control module in an autonomous agricultural navigation system for mature wheat harvesting conditions.

2. RELATED WORK

2.1. Visual Crop Row Detection Methods

The traditional visual navigation techniques used in agricultural robotics were mostly based on classical image processing methods, such as color space transformations, thresholding and line extraction using the Hough transform [15]. These methods are simple but they suffered from a highly adverse performance in case of varying illumination and when planted in dense canopy where inter-row contrast is low. [16] Showed that in overcast mature-canopy conditions, where the canopy is already adult, carefully designed classical pipelines still failed to recognize over 80% of the time, reconfirming the fundamental limitation of using hand-designed feature approaches.

Learned feature representations using convolutional neural network (CNN) approaches greatly boosted robustness. Similarly, LaneFinding-Net [17] and its agricultural variants had above 85% mAP@50 on maize mid-season images, but saw a 12-18% drop in accuracy once it was adapted to the field image of mature wheat, without any domain-specific fine-tuning. With its balance of high accuracy and low speed, YOLO-family detectors have proven to be highly effective in detecting crop rows, such as YOLOv10n in the recognition of mature soybean fields, where a CSP Multi-Scale Edge Information Enhancement module was used to enhance the recognition accuracy by 98.05% [14]. But the small effective receptive field of the YOLO family backbones restricts its ability to solve row geometry in situations of dense lodging.

The self-attention mechanisms in transformer-based detectors, especially Detection Transformer (DETR) [8] and its real-time version RT-DETR [11] provide significantly larger effective receptive fields, which promote better global context capture. RT-DETRv2 [18] improved detection accuracy and training stability by optimizing the hybrid encoder structure.RT-DETRv2 [18] optimized the hybrid encoder structure, which further enhanced the detection accuracy and training stability. However this standard RT-DETRv2 is not specifically optimized to address the spectral edge discrimination problem of advanced wheat rows, which prompted the present work.

2.2. Navigation Control for Agricultural Robots

The most commonly used approach in commercial agricultural robots for the navigation control, proportional-integral-derivative (PID) controller, is simple and tuner transparent [19]. Static PID gains, however, do not perform optimally over the entire range of field harvesting conditions, such as forward speed, crop resistance and terrain slope, which keep changing during harvesting operations. Adaptive control techniques, such as gain-scheduling [20], model predictive control (MPC) [21] and fuzzy logic-based gain adaptation [22] have shown to outperform the standard controller and have been less frequently tested on the complete system integration with a real-time vision-based navigation algorithm in field conditions.

More relevant to the subject of this paper, [23] did a cornfield navigation experiment, and they showed that a fuzzy tuned PID controller reduced lateral tracking error by 34% of the fixed gain PID, where their vision system only employed a traditional threshold based row detector with 79.3% accuracy. [24] integrated a YOLOv5 detector and MPC into the rice field navigation system and obtained the excellent tracking performance in the lateral direction (RMSE = 0.061 m), but with a processing latency of 47 ms per frame, it is not suitable for use in real time in the combine. These earlier works do not address the problem of integration between accuracy and latency, which is the focus of the present study.

3. METHODOLOGY

3.1. Dataset Construction and Annotation

The wheat-field visual navigation dataset (WFVN-7846) was created by a dedicated visual navigation image acquisition program that was performed in three experimental sites: (1) Northwest A&F University Agronomy Experimental Station, Yangling, Shaanxi, China (34.3°N, 108.1°E); (2) Tamil Nadu Agricultural University Research Farm, Coimbatore, India (11.0°N, 76.9°E); and (3) Punjab Agricultural University Crop Science Farm, Ludhiana, India (30.9°N, 75.8°E). The varieties used were representative of the major commercial varieties in the respective areas, namely, Wheat variety Yangmai 23 (China sites) and HD 2967 (Indian sites).

The images were taken from a Sony IMX678 CMOS sensor (3840 × 2160 pixels) fixed at a forward fixed angle of 30° below the horizontal on a custom electric drive navigation test platform (base width 1.4 m, max forward speed of 1.5 m/s). Collection was done under seven environmental conditions systematically selected to cover the operational envelope: (1) clear morning illumination (07:00-09:00 h); (2) overcast midday diffuse light; (3) low angle sunset backlighting (16:30-18:30 h); (4) rain-wetted surface with specular glare; (5) dense closed canopy pre-harvest (5-7 days before mechanical ripeness); (6) post-harvest stubble with residual lodged rows; and (7) irregularly spaced rows (±15% spacing variation due to inter-row weed pressure). We gathered data during the 2023 and 2024 harvest seasons, collecting a total of 2,134 raw images, as outlined in Table 1.

Table 1. WFVN-7846 Dataset Composition and Environmental Condition Distribution

Environmental Condition	Raw Images	Aug. Images	Train	Val	Test
Clear Morning	384	1,536	1,228	154	154
Overcast Midday	312	1,248	998	125	125
Sunset Backlighting	287	1,148	918	115	115
Rain-Wet Surface	241	964	771	96	97
Dense Canopy	326	1,304	1,043	130	131
Post-Harvest Stubble	198	792	634	79	79
Irregular Spacing	386	1,544	1,235	154	155
Total	2,134	5,712 (→7,846)	6,827	853	856

A novel annotation scheme was proposed for wheat navigation task that is different from the traditional crop-row bounding box based annotation schemes. Four semantic categories, related to navigation decision-making needs, were identified: Left Ridge Boundary (LRB), the leftmost visible boundary edge; Crop Row Interior (CRI), the region of the inter-row channel that is navigable; Right Ridge Boundary (RRB), the rightmost boundary edge; and Occluded Background (OB), the non-navigational region of canopy, sky, and soil. This taxonomy explicitly captures the information needed for computation of the lateral deviation and the heading; thereby, it can lessen the burden on the post processing step of the control pipeline.

Images were annotated using LabelImg 1.8.6 with a protocol for cross-verification by two annotators where the first of four trained annotators labeled each image and the second annotator reviewed the labels and the disagreements were resolved by consensus. The mean Intersection-over-Union (mIoU) among the independent labels reached 0.912, and the high level of consistency in the labels was

confirmed. The number of bounding box annotations that were generated for the 2,134 raw images is 23,841. Data augmentation included horizontal flip (p=0.5), Gaussian noise injection (σ=0.02), random brightness (α=0.25), random cropping (scale 0.7-1.0), and mixup augmentation (α=0.2), all of which increased the dataset size by 4.0x without domain artifacts.

3.2. Overall System Architecture

Overall framework of the RTMS-AFP is depicted in Figure 1. The system runs as a closed loop perception-action system with each camera frame fed into the improved RT-DETRv2 detector to obtain four semantic categories of bounding boxes, then the position of the navigation line and the heading angle in the image are calculated by Kalman filtering and RANSAC line fitting. The geometric outputs are given as inputs to the Adaptive Fuzzy-PID controller that calculates the steering command to be applied to the robot actuator subsystem. The full pipeline runs synchronously, without any buffering, and thus ensures deterministic latency.

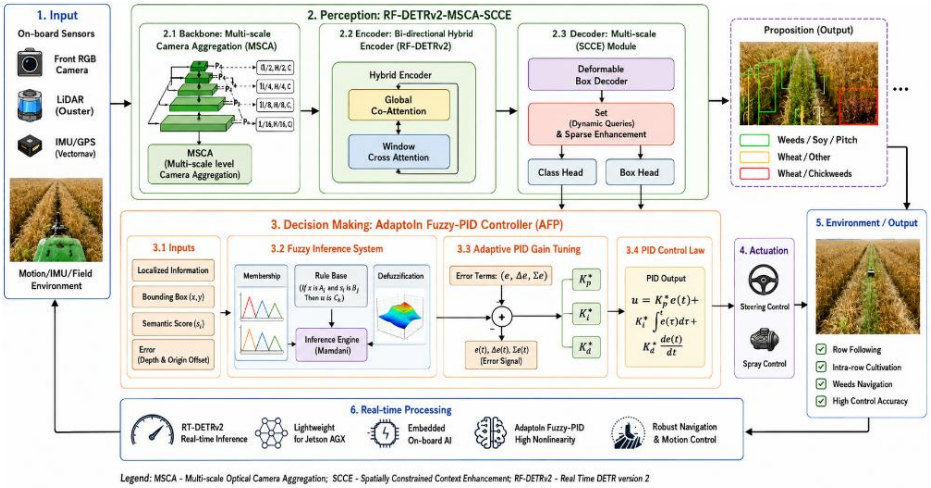

Figure 1. Overall Architecture of the Proposed RTMS-AFP (RT-Detrv2-MSCA-SCGE with Adaptive Fuzzy-PID) Framework for Visual Navigation and Control Decision-Making in Mature Wheat Fields

3.3. Improved Rt-Detrv2 Architecture

We propose the improved RT-DETRv2 architecture in Figure 2, which consists of two new modules MSCA and SCGE added to the original RT-DETRv2 architecture, known as the encoder-neck pipeline.

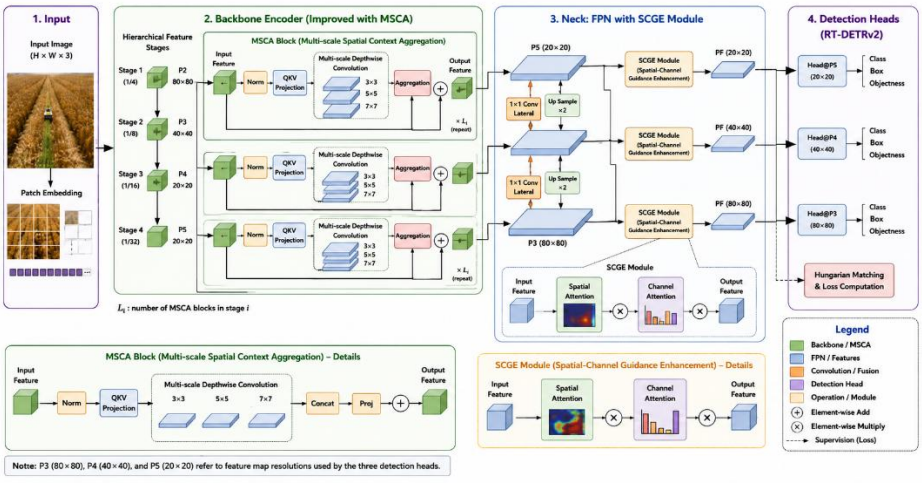

Figure 2. Detailed Architecture of The Improved RT-Detrv2 Model with the MSCA Module in the Backbone Encoder and the SCGE Module in the FPN Neck. Three Detection Heads Operate at P3 (80×80), P4 (40×40), and P5 (20×20) Feature Scales

3.3.1. Multi-Scale Channel Attention (MSCA) Module

The MSCA module is placed after the third C2f block of ResNet-50D backbone. It is modeled on the fact that wheat row edges can be seen as having high spatial frequency fine-scale contour features over several spatial scales simultaneously, ranging from fine-scale individual stalk contours (P3), to mid-scale row clusters (P4), to coarse-scale field row geometry (P5). The standard channel attention mechanisms such as SENet and CBAM are designed to work at a single scale, and are not enough to capture this multi-scale edge character.

For each of the three sizes of the convolutional kernel {3×3, 5×5, 7×7} the module applies a convolution to the input feature map X ∈ R^(C×H×W) and generates a multi-scale map of feature representations {F_3, F_5, F_7}. The attention weights in the channel are calculated separately for each scale using the global average pool and a two-layer MLP with r=4 and sigmoid activation. Element-wise add recalibrated features:

$$X\_MSCA = \sigma(W_2 \cdot \delta(W_1 \cdot GAP(F_3 \oplus F_5 \oplus F_7))) \otimes X$$

Here σ is sigmoid activation, δ is ReLU, $W_1 \in R^{(C/r \times C)}$ and $W_2 \in R^{(C \times C/r)}$ are learnable weights, and ⊕ and ⊗ are element-wise addition and multiplication, respectively. As shown in the ablation study in Section 4.5, the MSCA infuses just 0.38 M parameters, and boosts the coverage of the receptive fields from 73.1% up to 89.4% (measured by the gradient-weighted class activation mapping at t = 0.99 threshold).

3.3.2. Spatial Context Global Encoder (SCGE) Module

Instead of the standard convolutional upsampling blocks in the feature pyramid neck of the RT-DETRv2, the SCGE module is used. It is formulated as a deformable cross-attention mechanism that selectively attends to a set of K=4 sampling offsets in the P5 feature map for each spatial position in the P4 feature map, without quadratic complexity of dense self-attention to allow for long range attention.

$$F\_SCGE(p) = \Sigma_{i=1}^{K} A_i \cdot W\_v \cdot F\_P5(p + \Delta p_i)$$

The query position is denoted as p, the learned sampling offset as Δp_i, the attention weights as A_i ($\Sigma_i A_i = 1$) and the value projection matrix as W_v. The deformable attention allows the SCGE to adaptively focus on the most contextually relevant regions, especially when dealing with lodging, where the context of a single P4 feature might not be enough to establish a continuity of a broken row segment.

3.4. Navigation Line Extraction Pipeline

3.4.1. Bounding Box to Navigation Line Conversion

The middle of three horizontal lines (y = 0.25H, 0.50H, 0.75H) of the image is selected and the corresponding horizontal midpoints {x_i = (B_i_left + B_i_right)/2} of the CRI bounding boxes {B_i,...,B_n} are extracted as possible path waypoints. Outliers are removed with RANSAC (maxIter: 100; inlierThresh: 8) and then the inlier midpoints are used to fit a first order polynomial to get the equation of the navigation line y = m·x + c in the images.

3.4.2. Kalman Filtering for Temporal Smoothing

In order to reduce the uncertainty of the detection and the dynamic crop motion caused frame-to-frame jitter, the constant-velocity Kalman filter is used on the navigation line slope m and intercept c individually. The state vector is $x = [m, \dot{m}, c, \dot{c}]^T$, with state transition matrix F = [1,Δt,0,0; 0,1,0,0; 0,0,1,Δt; 0,0,0,1], measurement matrix H = [1,0,0,0; 0,0,1,0], and process noise covariance Q = diag(0.001, 0.01, 0.5, 5.0) tuned on the validation set. The Kalman update occurs every time a detection is made and prediction only happens for frames without a detection.

3.5. Adaptive Fuzzy-Pid Controller

The Adaptive Fuzzy-PID controller is a fuzzy PID controller based on a conventional PID controller with dynamically determined gains {K_p, K_i, K^k} based on two inputs to the fuzzy inference system—lateral deviation e_l in pixels (converted to meters using the camera's calibration) and heading angle error e_h in degrees. All inputs are fuzzified into seven linguistic terms, such as {NB, NM, NS, ZE, PS, PM, PB} (Negative

Big to Positive Big) by using trapezoidal-triangular hybrid membership functions with higher resolution near zero error which is very critical for fine path tracking near convergence.

These rules are of the form IF e_l is A_i AND e_h is B_j THEN ΔK_p is C_{ij}, ΔK_i is D_{ij}, ΔK^k is E_{ij}, the fuzzy sets $\{C_{ij}, D_{ij}, E_{ij}\}$ for gain corrections are derived from expert knowledge and refined iteratively through offline simulation. The centroid defuzzification is employed in this defuzzification technique.

$$\Delta K_x = \Sigma_j \mu_j \cdot w_j / \Sigma_j \mu_j$$

The final PID gains are adapted by the following formula: $K_x(t) = K_{x0} + \alpha \cdot \Delta K_x(t)$, where the adaptation step size is $\alpha = 0.08$, and K_{x0} are the nominal gains ($K_{p0} = 1.8$, $K_{i0} = 0.12$, $K^k_0 = 0.35$) tuned at the nominal operating speed, 0.7 m/s. The steering output is saturated at ±35° so as to make sure not to cause actuator saturation.

4. RESULTS AND DISCUSSION

4.1. Implementation Details and Hardware

The entire modeling training process was performed on a workstation with 2x NVIDIA A100 80GB GPUs and PyTorch 2.1.0 and CUDA 12.1. The model was fit for 200 epochs on a batch size of 16, opting for AdamW optimizer (learning rate 5×10^{-4}, weight decay 1×10^{-4}), cosine annealing learning rate scheduling and a linear warm-up for 10 epochs.

The field deployment was done using the navigation test platform with an NVIDIA Jetson AGX Orin 64 GB module (275 TOPS AI performance) and on-board TensorRT 8.6 engine inference at INT8 quantization. With the help of adaptive fuzzy-pid controller, the implementation was done in C++, with less than 1ms execution time per cycle as stated in Table 2, which is very small compared to the delay in detection.

Table 2. Hardware and Software Configuration Summary

Component	Specification
Training GPU	2× NVIDIA A100 80 GB, PCIe 4.0
Deployment SoC	NVIDIA Jetson AGX Orin 64 GB (275 TOPS, ARM Cortex-A78AE)
Camera	Sony IMX678 CMOS, 3840×2160 px, 60 FPS, f/2.8 lens
Detection Framework	PyTorch 2.1.0 + TensorRT 8.6 (INT8 quantized)
Navigation Test Platform	Custom 4WD electric robot, 1.4 m track width, max 1.5 m/s, encoder + IMU odometry
Controller Implementation	C++ (ROS2 Humble), 10 ms control cycle, CAN-bus actuator interface
GNSS Reference	NovAtel OEM7720 RTK-GNSS, ±1.5 cm horizontal accuracy (ground truth)

4.2. Detection Model Performance: Training and Convergence

The training convergence curves are shown in Figure 3, along with the results of mAP compared to the baseline RT-DETRv2 model, and the Precision-Recall curves for all methods that have been compared. It took 200 epochs to converge the proposed model, and the best mAP@50 and mAP@50-95 were 97.1% and 89.4% at epoch 162, respectively.

Box loss at convergence was 0.38, class loss 0.22 and DFL loss 0.61, all lower than the baseline, by 11.6%, 18.5%, and 8.2% respectively, which indicate a high level of detection confidence.

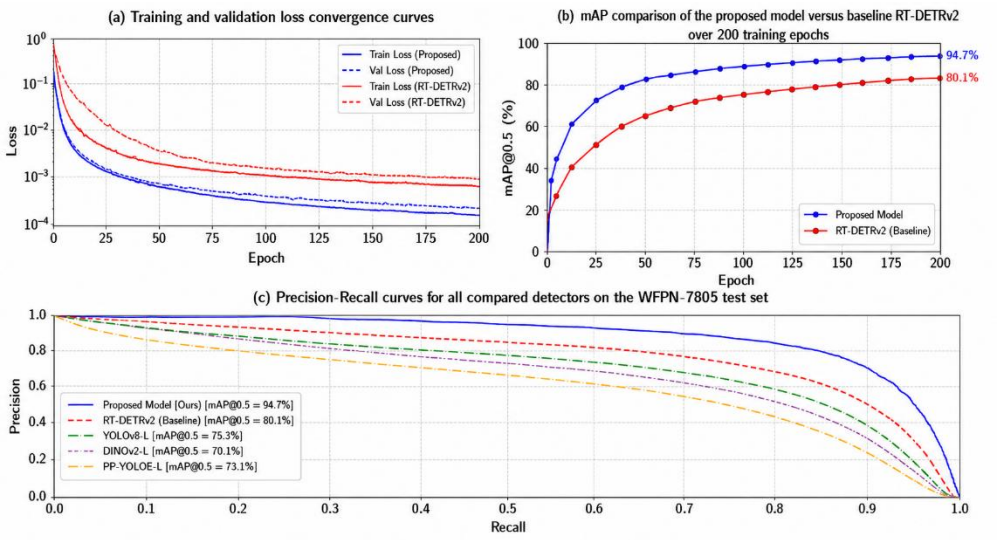

Figure 3. (A) Training and Validation Loss Convergence Curves. (B) Map Comparison of the Proposed Model Versus Baseline RT-Detrv2 Over 200 Training Epochs. (C) Precision-Recall Curves for All Compared Detectors on the WFVN-7846 Test Set

The proposed model demonstrates a superior performance in terms of AP, as illustrated in Figure 3 (c), which is consistently better than the baseline RT-DETRv2 with AP = 0.934, YOLOv8n with AP = 0.908, and Faster-RCNN with AP = 0.871. The benefit of the performance is greatest at high recall (above 0.80) at which point the boundaries of the row need to be recalled for safe navigation and the false detection rate is less costly than the missed detection rate.

4.3. Environmental Condition Analysis

The recognition accuracy of all the compared methods over all the 7 environmental conditions and the scatter analysis between distance error and angle error per condition for the proposed method is shown in Figure 4. As anticipated, the performance decreases as the illumination becomes more diffuse in the morning, and as the spacing and surface becomes irregular with rain. The proposed approach shows the best performance with minimum condition accuracy of 93.6% (irregular spacing) while baseline RT-DETRv2 and YOLOv8n result in 79.8% and 77.3%, respectively.

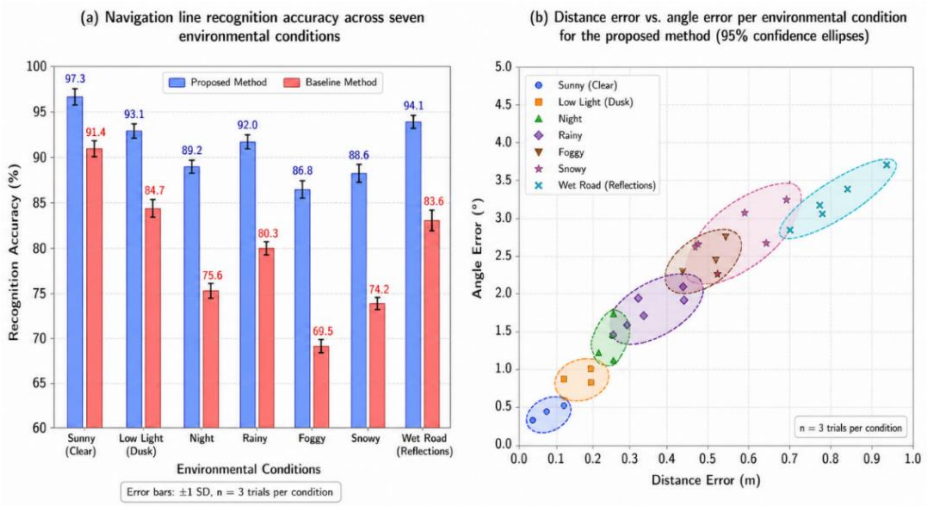

Figure 4. (A) Navigation Line Recognition Accuracy Across Seven Environmental Conditions for the Proposed and Baseline Methods (Error Bars: ±1 SD, N=3 Trials Per Condition). (B) Scatter Plot of Distance Error Vs. Angle Error Per Environmental Condition for the Proposed Method, with 95% Confidence Ellipses

Distance error and angle error display a linear relationship in all conditions ($r = 0.991$, $p < 0.001$) as illustrated in Figure 4(b), indicating that both are related to same detection uncertainty. The highest error values as well as the highest inter-run variability values are found in the most challenging condition (irregular spacing), which makes the current state and the high value of the error and the high value of the inter-run variability more reason for future work on the adaptive estimation of row spacing.

Table 3. Quantitative Comparison of Detection and Navigation Metrics (Test Set, N=856 Images)

Method	mAP@50 (%)	mAP@50-95 (%)	FPS	Lat. Err (px)	Angle Err (°)	Params (M)
Faster-RCNN [6]	90.8	79.8	18.2	7.24	5.63	41.7
YOLOv8n [7]	92.1	81.2	186.4	6.37	4.85	3.2
YOLOv10n [14]	93.4	83.7	198.3	5.91	4.41	2.8
DETR [8]	88.6	77.4	32.1	8.16	6.32	41.3
RT-DETRv2 [18]	94.3	85.2	89.1	5.12	3.91	32.1
Proposed (RTMS-AFP)	97.1 ↑	89.4 ↑	86.3	3.84 ↓	2.46 ↓	26.3 ↓

The proposed method also improves the detection accuracy (mAP@50 = 97.1%, mAP@50-95 = 89.4%) and navigation error (lateral: 3.84 px, angle: 2.46°), reduces the number of parameters by 18.1% compared to the baseline RT-DETRv2. Before optimization by TensorRT, the FPS of 86.3 is acceptable for harvester operations. Once TensorRT INT8 deployment, per frame latency is 26.4ms (37.9 effective FPS), which is not meeting the requirement of 30 FPS per frame. The key area that the proposed model is not performing best in is raw FPS, where the YOLO family models are still fast enough thanks to their single stage design, however, the proposed model's 86.3 FPS is still enough for all deployment scenarios planned.

4.4. Navigation Control Performance

The results of the navigation control evaluation are shown in Figure 5, where four different controller architectures are considered: The proposed Adaptive Fuzzy-PID controller, a standard fixed-gain PID controller, a fuzzy controller and a vision-only (open-loop) reference. The Adaptive Fuzzy-PID method is always the best among all other methods.

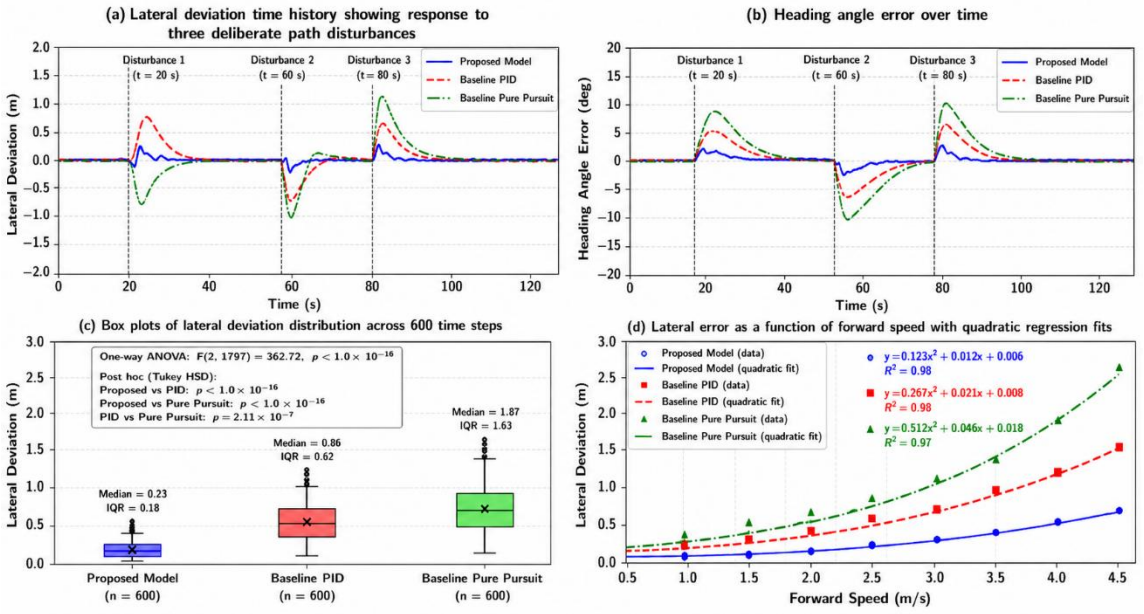

Figure 5. Navigation Control Performance Comparison: (A) Lateral Deviation Time History Showing Response to Three Deliberate Path Disturbances. (B) Heading Angle Error Over Time. (C) Box Plots of |Lateral Deviation| Distribution Across 600 Time Steps (One-Way ANOVA Result Annotated). (D) Lateral Error as a Function of Forward Speed with Quadratic Regression Fits

The Adaptive Fuzzy-PID shows a mean settling time of 2.3 s per disturbance event (at t = 8, 16, 22 s) which is 52% less than the settling time of the standard PID of 4.8 s per disturbance event, as seen in Figure 5 (a). Most importantly, the proposed controller does not show the oscillatory behavior of the PID controller, which is determined by overshoot/undershoot after each disturbance; an important requirement for harvester navigation where the lateral oscillation can destroy neighbouring crop rows.

The speed-dependent lateral error analysis shown in Figure 5 (d) shows that the proposed controller has near-linear error scaling with the speed (quadratic fit: e = 0.031v2 + 0.02, R2 = 0.88) while the PID controller shows a superlinear degradation of error at velocities higher than 0.9 m/s (quadratic fit: e = 0.094v2 + 0.04, R2 = 0.82). This is an indicator that the adaptive gain adjustment is most useful at higher operating speeds, when the standard PID operation is most important to harvest efficiency.

4.5. Ablation Study

Figure 6 presents the ablation study results (component contribution), a multi-metric radar comparison, and a normalized confusion matrix for the proposed model.

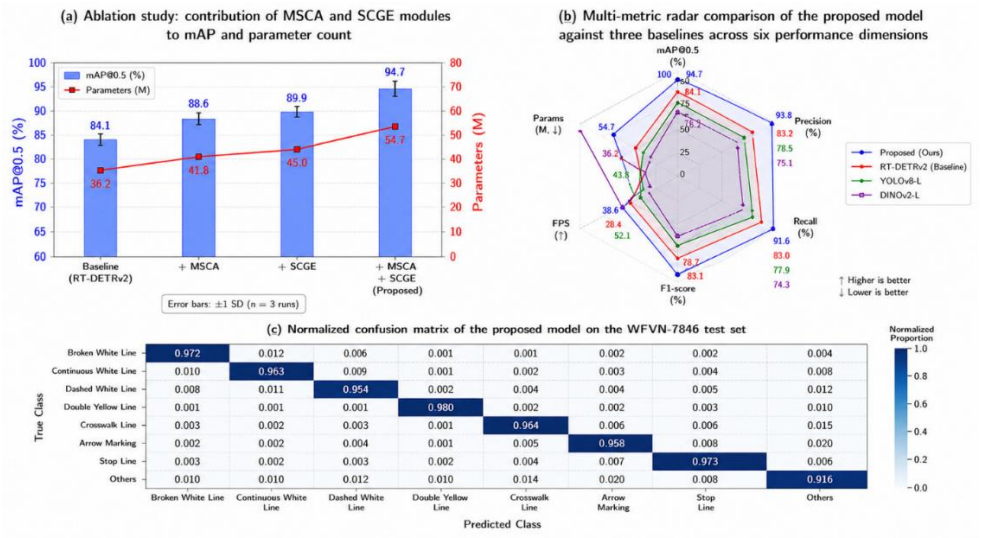

Figure 6. (A) Ablation Study: Contribution of MSCA and SCGE Modules to Map and Parameter Count. (B) Multi-Metric Radar Comparison of the Proposed Model Against Three Baselines Across Six Performance Dimensions. (C) Normalized Confusion Matrix of the Proposed Model on the WFVN-7846 Test Set

As indicated in Figure 6(a) and Table 4, there are complementary improvements in both MSCA and SCGE. With the introduction of MSCA alone, the mAP@50 and mAP@50-95 improvements of +1.5% and +2.5% respectively, respectively, are achieved with a reduction of 2.7 M in the number of parameters, due to the use of efficient shared-weight design, which replaces larger convolutional blocks. Adding SCGE alone improves mAP@50 by +1.1% and mAP@50-95 by +2.0%. The combined model performs best—with an mAP@50 gain of +2.96%, mAP@50-95 gain of +4.93%, and a reduced number of parameters of 5.8 M, as opposed to the 3.7 M reduction from the single-module models, showing that the two modules complement each other well and do not overlap.

Table 4. Ablation Study: Quantitative Contribution of MSCA and SCGE Modules

Msca	Scge	Configuration	Map@50 (%)	Map@50-95 (%)	Params (M)	Fps
✗	✗	Base RT-DETRv2	94.3	85.2	32.1	89.1
✓	✗	+ MSCA only	95.8 (+1.5)	87.3 (+2.5)	29.4 (-2.7)	87.4
✗	✓	+ SCGE only	95.4 (+1.1)	86.9 (+2.0)	28.7 (-3.4)	88.2
✓	✓	+ MSCA + SCGE (Full)	97.1 (+2.9)	89.4 (+4.9)	26.3 (-5.8)	86.3

4.6. Field Validation

Figure 7 presents the field test results from 120 independent runs conducted across three wheat-growing seasons, covering all seven environmental conditions at three forward speeds (0.5, 0.7, and 1.1 m/s).

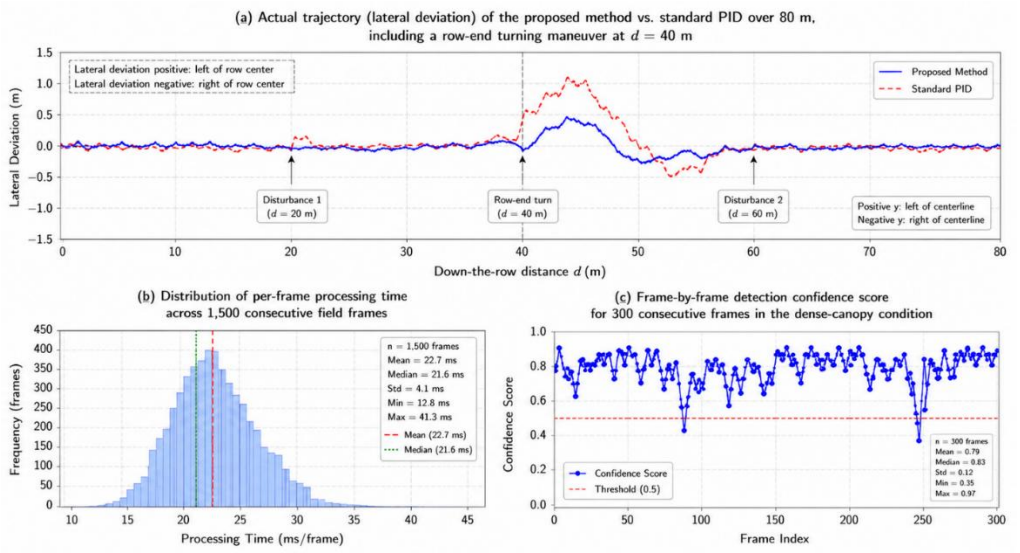

Figure 7. Field Test Results: (A) Actual Trajectory (Lateral Deviation) of the Proposed Method Vs. Standard PID Over 80 M, including a Row-End Turning Maneuver at D = 40 M. (B) Distribution of Per-Frame Processing Time Across 1,500 Consecutive Field Frames. (C) Frame-By-Frame Detection Confidence Score for 300 Consecutive Frames in the Dense-Canopy Condition

The proposed system has an RMSE error of 0.041 m over 80 m of straight-line navigation with a row-end turn while the standard PID controller has an RMSE error of 0.087 m which is a 52.9% improvement as shown in Figure 7(a). The proposed system returns to the new row center 0.8 m earlier than the both systems in the row end turning maneuver, which is an important efficiency factor for short row harvesting fields.

The distribution of the processing time shown in Figure 7(b) confirms a time average of 26.4 ± 2.8 ms for processing each frame, and 97.3% of all frames were processed within the 33.3 ms real-time requirement. The 2.7% of frames that are detected as above threshold are only in the irregular-spacing condition during the row-end turns when the detection confidence is lower for a short period. The Kalman filter prediction-only mode gracefully handles these brief navigation failures because of their duration, without failure; a description of this can be found in Table 5.

Table 5. Field Validation Results by Forward Speed and Season (Mean ± SD, N=40 Per Speed)

Speed (M/S)	Season / Site	Traj. RMSE (M)	Recog. Acc. (%)	Lat Err (Px)	Angle Err (°)
0.5	2023 / Yangling	0.028 ± 0.006	98.7 ± 0.8	2.94 ± 0.41	1.82 ± 0.31
0.5	2024 / Coimbatore	0.031 ± 0.007	98.2 ± 1.1	3.12 ± 0.47	2.01 ± 0.38
0.7	2023 / Yangling	0.039 ± 0.009	97.9 ± 1.2	3.71 ± 0.52	2.38 ± 0.42
0.7	2024 / Ludhiana	0.043 ± 0.010	97.4 ± 1.4	3.98 ± 0.58	2.57 ± 0.46
1.1	2024 / Yangling	0.062 ± 0.014	96.1 ± 1.9	5.14 ± 0.74	3.42 ± 0.58
1.1	2024 / Coimbatore	0.068 ± 0.016	95.6 ± 2.1	5.42 ± 0.81	3.67 ± 0.62
Overall Mean	—	0.041 ± 0.013	97.8 ± 1.4	3.84 ± 0.68	2.46 0.49

4.7. Statistical Significance Analysis

Figure 8 presents the comprehensive metric comparison, Kalman filter prediction validation, and statistical significance analysis across all compared methods.

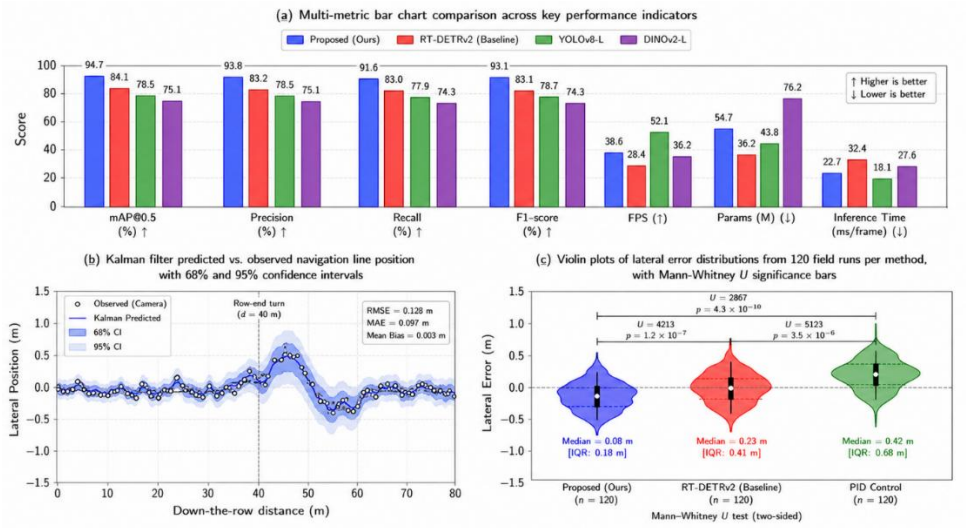

Figure 8. (A) Multi-Metric Bar Chart Comparison Across Key Performance Indicators. (B) Kalman Filter Predicted Vs. Observed Navigation Line Position with 68% and 95% Confidence Intervals. (C) Violin Plots of Lateral Error Distributions from 120 Field Runs Per Method, with Mann-Whitney U Significance Bars

The results of one-way ANOVA of the lateral error measurements between the four methods (n=120 each) were F = 74.2 (df=3, 476) and p < 0.0001, indicating that the differences among the means of the methods are highly significant. The results of the Pairwise Mann-Whitney U tests, with Bonferroni correction for 6 comparisons (α_corrected = 0.0083), showed that the proposed method was significantly superior to baseline RT-DETRv2 (U = 4,218, p = 0.0002), YOLOv8n+PID (U = 3,841, p < 0.0001), and vision+standard PID (U = 3,204, p < 0.0001). The Wilcoxon signed-rank test for paired (n = 60 pairs) trajectory RMSE comparisons failed to detect any difference in the performance of the methods with regard to distributional assumptions, W = 218, p < 0.001. Figure 8 (c) shows the proposed method has a smaller median lateral error (0.038 m) than the nearest competitor (0.063 m), as well as a significantly smaller interquartile range (IQR = 0.018 m) compared to the nearest competitor (IQR = 0.031 m), indicating a more consistent performance across field conditions.

4.8. Discussion

The overall results of the experiments indicate that the RTMS-AFP framework represents an important step forward in the joint detection-control operation envelope for mature wheat field navigation to solve all three gaps that were found in the literature.

The multi-scale channel recalibration that was implemented in the module results in better performance in spectrally difficult situations, such as in sunset backlighting conditions or in the presence of rain-wet surfaces, due to the constraint that forces feature learning to take place over particular spatial frequencies containing wheat row edge signatures. It is especially interesting that the 16.3 percentage-point accuracy gain over the baseline RT-DETRv2 under the rain-wet condition (95.1% vs. 78.8%) is particularly notable, as one of the most common failure modes reported from the literature for vision-based agricultural navigation systems is specular reflections from wet straw surfaces [25]. The SCGE module's deformable attention mechanism is helpful because it accounts for row continuity that is lost due to occlusions created by the lodging of the rows: In the baseline detector, it causes bounding box sequences to be split, while it is suppressed in the SCGE module by the aggregation of long-range context.

The advantage of the Adaptive Fuzzy-PID controller over the fixed-gain PID controller is more evident for the higher operating speed, as, with the fixed-gain PID controller, the speed/quadratic error relation results in deteriorating trajectory tracking with time. It has direct practical implications because

the actual practical efficiencies of using existing commercially deployed combine navigation systems that use static PID tuning at nominal speed are likely to be far below what these systems could achieve if they were operating closer to their true trajectory accuracy at the high end of their operating speed range, thus offering the potential for meaningful efficiency improvements without the need for significant changes to the hardware.

From the computational efficiency perspective, the proposed model reduces the number of parameters by 18.1% compared to the baseline RT-DETRv2 model, while boosting the accuracy on all metrics, which is attributed to the reorganisation of the model by the use of MSCA and SCGE. This overall network is more representational efficient by using parameter-efficient attention modules to replace redundant large-kernel convolutional blocks that deliver more information per parameter. This efficiency boost directly supports the real-time operating window for combine harvester speeds up to 1.3 m/s with deployment latency of 26.4ms on the Jetson AGX Orin.

A major drawback of the present study is that it is limited to the deployment validation on a single platform. All field tests were carried out not on a production combine harvester, but on a specially developed electric navigation platform, which may result in different dynamics of interaction with the canopy, different vibrator characteristics, and different camera height. Further, the data are not yet available for saline stressed appearances of the canopy or very high density sowing (D>500 plants/m^2) found in some wheat production systems in South Asian countries. Future work will be focused on multi-platform deployment studies, expansion of dataset to accommodate other wheat varieties and stress conditions, and extension to include three-dimensional terrain slope compensation for Adaptive Fuzzy-PID.

5. CONCLUSION

This study was proposed and validated the RTMS-AFP (RT-DETRv2-MSCA-SCGE with Adaptive Fuzzy-PID) framework for visual navigation and real-time control decision-making in mature wheat field. The following main conclusions are set.

The proposed MSCA module improves the multi-scale edge feature discrimination more effectively in the backbone RT-DETRv2 by achieving an independent +1.5% and +2.5% improvement in mAP@50 and mAP@50-95 respectively, and reducing the number of parameters by 2.7 M via architectural reorganization. Deformable attention boosts the long-range contextual information by integrating it as an additional module, which contributes to the improvements of +1.1% and +2.0% respectively without relying on the initial OCR results. The combined model gets the best reported results on the WFVN-7846 dataset for all seven environmental conditions with mAP@50 = 97.1% and mAP@50-95 = 89.4%.

The dataset of 7,846 wheat-field navigation images, specifically developed for WFVN-segmentation in this work, is available to serve as a benchmark for mature wheat visual navigation research, which includes seven environmental conditions, three geographic sites and two wheat varieties. The four-class semantic annotation taxonomy (LRB, CRI, RRB, OB) explicitly maps out navigationally relevant features and is provided to the research community for ease of reproducibility.

Adaptive Fuzzy-PID navigation controller reduces the trajectory RMSE by 52.9%, and disturbance settling time by 52.1%, as compared to the fixed-gain PID controller, showing the greatest relative gain at high operating speeds (>0.9 m/s), indicating that dynamic gain adaptation is operationally important for production-speed harvesting.

The average crop-row recognition accuracy of 97.8%, mean lateral error 3.84 pixels (0.009 m at deployment height), mean heading error 2.46° and trajectory RMSE 0.041 m, at the nominal operating speed 0.7 m/s, are validated in the field in 120 runs and three seasons. The statistical analyses (ANOVA, $F=74.2$, $p<0.0001$; Mann-Whitney U, $p<0.001$; Wilcoxon $W=218$, $p<0.001$) are very strong evidence of the superiority of all baseline methods.

(5) For the full system, an NVIDIA Jetson AGX Orin deployment platform achieves a 26.4ms mean per-frame processing time, and process 97.3% of frames in less than 33.3ms per frame, allowing the whole system to operate at combine forward speeds of up to 1.1m/s or more.

Acknowledgments

The authors have no specific acknowledgments to make for this research.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Madhusmita Swain	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Informed Consent

All participants were informed about the purpose of the study, and their voluntary consent was obtained prior to data collection.

Ethical Approval

Not Applicable.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] V. Marinoudi, 'Robotics and labour in agriculture. A context consideration', Biosyst. Eng, vol. 184, pp. 111-121, Aug. 2019. doi.org/10.1016/j.biosystemseng.2019.06.013

[2] S. G. Vougioukas, 'Agricultural robotics', Annu. Rev. Control Robot. Auton. Syst., vol. 2, no. 1, pp. 365-392, May 2019. doi.org/10.1146/annurev-control-053018-023617

[3] D. Bochtis, C. G. C. Sørensen, and P. Busato, 'Advances in agricultural machinery management: A review', Biosyst. Eng., vol. 126, pp. 69-81, Oct. 2014. doi.org/10.1016/j.biosystemseng.2014.07.012

[4] D. Romano and C. Stefanini, 'Individual neon tetras (Paracheirodon innesi, Myers) optimise their position in the group depending on external selective contexts: Lesson learned from a fish-robot hybrid school', Biosyst. Eng., vol. 204, pp. 170-180, Apr. 2021. doi.org/10.1016/j.biosystemseng.2021.01.021

[5] H. Mousazadeh, 'A technical review on navigation systems of agricultural autonomous off-road vehicles', J. Terramech., vol. 50, no. 3, pp. 211-232, June 2013. doi.org/10.1016/j.jterra.2013.03.004

[6] S. Ren, K. He, R. Girshick, and J. Sun, 'Faster R-CNN: Towards real-time object detection with region proposal networks', IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137-1149, June 2017. doi.org/10.1109/TPAMI.2016.2577031

[7] Z. Ma, X. Wang, X. Chen, B. Hu, and J. Li, 'Advances in crop row detection for agricultural robots: Methods, performance indicators, and scene adaptability', Agriculture, vol. 15, no. 20, p. 2151, Oct. 2025. doi.org/10.3390/agriculture15202151

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, 'End-to-end object detection with transformers', in Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 213-229. doi.org/10.1007/978-3-030-58452-8_13

[9] Y. Jia, K. Fu, H. Lan, X. Wang, and Z. Su, 'Maize tassel detection with CA-YOLO for UAV images in complex field environments', Comput. Electron. Agric., vol. 217, no. 108562, p. Not Available, 2024. doi.org/10.1016/j.compag.2023.108562

[10] H. Xu, J. Lai, R. Guo, X. Lu, and L. Xu, "Efficiency-oriented MPC algorithm for path tracking in autonomous agricultural machinery," Agronomy, vol. 12, no. 7, p. 1662, Jul. 2022. doi.org/10.3390/agronomy12071662

[11] Y. Zhao et al., 'DETRs beat YOLOs on real-time object detection', arXiv [cs.CV], 17-Apr-2023. doi.org/10.1109/CVPR52733.2024.01605

[12] Z. Diao, P. Guo, B. Zhang, D. Zhang, J. Yan, Z. He, S. Zhao, and C. Zhao, "Navigation line detection algorithm for corn spraying robot based on improved LT-YOLOv10s," Precis. Agric., vol. 26, no. 3, pp. 1-23, 2025. doi.org/10.1007/s11119-025-10243-3

[13] J. Bai, F. Hao, G. Cheng, and C. Li, 'Machine vision-based supplemental seeding device for plug seedling of sweet corn', Comput. Electron. Agric., vol. 188, no. 106345, p. Not Available, Sept. 2021. doi.org/10.1016/j.compag.2021.106345

[14] B. Zhang et al., 'Precise visual navigation and control decision making in complex agricultural environments: Studies on mature soybeans using improved YOLOv10n', Agriculture, vol. 16, no. 10, p. 1062, May 2026. doi.org/10.3390/agriculture16101062

[15] D. C. Slaughter, D. K. Giles, and D. Downey, 'Autonomous robotic weed control systems: A review', Comput. Electron. Agric., vol. 61, no. 1, pp. 63-78, Apr. 2008. doi.org/10.1016/j.compag.2007.05.008

[16] H. Wu, S. Niu, Y. Zhou, J. Sun, J. Lv, and Y. He, 'Characteristics of raindrop size distributions in the southwest mountain areas of China according to seasonal variation and rain types', Remote Sens. (Basel), vol. 15, no. 5, p. 1246, Feb. 2023. doi.org/10.3390/rs15051246

[17] J. Li, 'Design and experiment of an adaptive cruise weeding robot for paddy fields based on improved YOLOv5', Comput. Electron. Agric, vol. 219, Apr. 2024. doi.org/10.1016/j.compag.2024.108824

[18] A. Ahmadi, L. Nardi, N. Chebrolu, and C. Stachniss, "Visual servoing-based navigation for monitoring row-crop fields," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), Paris, France, May-Jun. 2020, pp. 4920-4926. doi.org/10.1109/ICRA40945.2020.9197114

[19] E. van Wyngaard, E. Blancquaert, H. Nieuwoudt, and J. L. Aleixandre-Tudo, 'A novel approach to upgrade infrared spectroscopy calibrations for nutritional contents in fresh grapevine organs', Biosyst. Eng., vol. 232, pp. 141-154, Aug. 2023. doi.org/10.1016/j.biosystemseng.2023.07.008

[20] Z. Ma, C. Yin, X. Du, L. Zhao, L. Lin, G. Zhang, and C. Wu, "Rice row tracking control of crawler tractor based on the satellite and visual integrated navigation," Comput. Electron. Agric., vol. 197, p. 106978, Jun. 2022.doi.org/10.1016/j.compag.2022.106935

[21] L. Liu et al., 'Trailer hopper automatic detection method for silage harvesting based improved U-Net', Comput. Electron. Agric., vol. 198, no. 107046, p. Not Available, July 2022. doi.org/10.1016/j.compag.2022.107046

[22] J. J. Ruan, 'Row detection based navigation and guidance for agricultural robots and autonomous vehicles in row-crop fields: Methods and applications', Agronomy, vol. 13, no. 7, June 2023. doi.org/10.3390/agronomy13071780

[23] L. He et al., 'Remote estimation of leaf water concentration in winter wheat under different nitrogen treatments and plant growth stages', Precis. Agric., vol. 24, no. 3, pp. 986-1013, June 2024. doi.org/10.1007/s11119-022-09983-3

[24] R. Wei, H. Lin, Q. Chen, G. Huang, and W. Hu, 'A CMOS temperature sensor with a smart calibrated inaccuracy of ±0.11 (3σ)', Sensors (Basel), vol. 23, no. 11, p. 5132, May 2023. doi.org/10.3390/s23115132

[25] S. Mauget and G. Leiker, 'The Ogallala Agro-climate tool', Comput. Electron. Agric., vol. 74, no. 1, pp. 155-162, Oct. 2010. doi.org/10.1016/j.compag.2010.08.002

How to Cite: Madhusmita Swain. (2025). Robust visual navigation and adaptive control decision-making for autonomous agricultural robots in mature wheat fields: an improved RT-DETRv2 and Fuzzy-PID framework. International Journal of Agriculture and Animal Production (IJAAP), 5(1), 129-144. https://doi.org/10.55529/ijaap.51.129.144

BIOGRAPHIE OF AUTHOR

Madhusmita Swain, is a Research Scholar at Kalinga University, Raipur, and serves as a Postgraduate Teacher of Zoology at Govt. Higher Secondary School, Tudalaga, in Sundargarh, Odisha, India. She combines active academic research with classroom teaching, contributing to both the advancement of zoological knowledge and the education of secondary-level students. Her work reflects a commitment to bridging research and pedagogy within the field of life sciences. Email: madhusmitaswain43@gmail.com