NHL STENDEN
computer vision & data science

Salient Region-of-Interest Detection in Multispectral Agricultural Imagery

NHL Stenden Lectoraat in Computer Vision & Data Science

Çağkan Gürsoy
Supervisors: Willem Dijkstra

Abstract—Modern agriculture faces growing challenges in maintaining yield and sustainability under climate stress and resource constraints. Identifying fine-grained deviations in crop growth patterns at scale is critical for supporting localized interventions, yet current field inspection methods remain labor-intensive and inconsistent. Drone-based multispectral imaging offers a promising alternative by providing detailed views of plant health and field structure, but automated detection of subtle variations in such imagery remains an open problem. This study presents a transformer-based object detection framework, adapted from Detection Transformer (DETR), for analyzing high-resolution agricultural drone imagery. The architecture integrates spatially preserving tiling to maintain fine visual detail, precise RGB-NIR alignment, and multiple spectrum fusion strategies to explore how spectral inputs contribute to detection. Model behavior is further analyzed through Grad-CAM and attention map visualizations to provide insight into spatial reasoning and spectral focus. Models based on RGB inputs consistently achieve the most reliable performance, with the green channel contributing strongly to discriminative features. NIR input adds limited value, even when integrated through mid-level fusion. High-performing models focus attention on plant morphology and structured row patterns, while less effective configurations exhibit scattered or background-focused attention. Maintaining spatial resolution is essential, as performance degrades significantly with excessive downscaling. These findings inform the design of scalable detection pipelines for precision agriculture, emphasizing the importance of visible-spectrum cues and spatial detail while guiding future research toward more effective spectral fusion techniques.

Index Terms—Precision Agriculture, Object Detection, Transformer Models, Multispectral Imaging, Drone Imagery, RGB-NIR Fusion, Saliency Analysis, Attention Visualization, High-Resolution Remote Sensing, Crop Monitoring

1 Introduction

Agriculture remains central to global food security but faces increasing stress from climate change, soil degradation, and resource scarcity [1]. With rising population demands and environmental constraints, there is a growing need for sustainable agricultural practices that improve yield efficiency while reducing ecological impact [2, 3].

Precision agriculture has emerged as a key approach to address these challenges by using sensor technologies, remote sensing, and data-driven analytics to monitor crop conditions and enable timely, localized interventions [4, 5]. A critical aspect of this approach is the ability to detect anomalies at the field level, including missing and underdeveloped crops, which serve as early indicators of localized stress, equipment failure, or environmental variability.

However, achieving effective and scalable anomaly detection in large agricultural fields remains a significant challenge. While traditional field inspections remain widely practiced, they are labor-intensive, time-consuming, and their reliability is often questionable due to subjectivity and limited spatial coverage [6]. As a scalable alternative, drone-based multispectral imaging provides high-resolution views of crop health across entire fields. In particular, RGB and Near-Infrared (NIR) imaging capture complementary cues: RGB highlights structural features, while NIR emphasizes vegetation vigor. Together, these modalities enhance the ability to detect subtle plant anomalies [7, 8]. Yet, transforming these rich multispectral images into actionable insights remains a complex task that requires robust interpretation techniques.

Recent advances in deep learning have enabled the automated interpretation of aerial and multispectral imagery. Among these methods, Salient Object Detection (SOD) has gained prominence due to its ability to highlight regions of interest (ROIs) by emulating human visual attention and suppressing background noise [9]. In agricultural applications, existing research has mostly focused on plant-level detection of visible anomalies, such as pest damage or general signs of crop stress. In contrast, field-level detection of missing plants and fine-grained identification of underdeveloped plants remains underexplored. These tasks present distinct challenges for automated analysis. Missing plants offer no direct visual signal and must be inferred from disruptions in planting geometry, while underdeveloped plants are visually similar to healthy ones and require sensitivity to subtle morphological differences.

Despite the potential of deep learning and multispectral imaging in agriculture, several design challenges remain insufficiently addressed. First, the contribution of Near-Infrared (NIR) imagery to detecting subtle anomalies such as underdeveloped crops is unclear, particularly within transformer-based architectures. Second, the minimum image resolution required to maintain detection reliability is not well defined, which poses challenges for optimizing drone flight altitude, image capture settings, and processing overhead. Finally, there is limited work on systematically evaluating how these factors affect the model's ability to detect field-level anomalies, especially in distinguishing missing plants and fine-grained underdevelopment from healthy vegetation.

To address these gaps, this study proposes a transformer-based detection framework tailored for structured agricultural fields using high-resolution multispectral drone imagery. The approach builds on the Detection Transformer (DETR) and introduces several domain-specific enhancements, including spatially preserving tiling for large-format inputs, RGB-NIR fusion through early and mid-level integration strategies, and attention-based interpretability tools. The framework is evaluated through a systematic analysis of image resolution by controlled downsampling and through comparative testing of multispectral fusion strategies. Specifically, we implement and compare early fusion (channel-wise concatenation of RGB and NIR) and mid-level fusion (separate RGB and NIR backbones with

- *Çağkan Gürsoy is a Computer Vision & Data Science Master's student at the NHL Stenden University of Applied Sciences, E-mail: cagkan.gursoy@student.nhlstenden.com.*
- *Willem Dijkstra is a researcher at the NHL Stenden Lectoraat in Computer Vision & Data Science, E-mail: willem.dijkstra@nhlstenden.com.*

mid-level feature fusion) within a DETR-based architecture, assessing detection accuracy to determine how NIR contributes to identifying anomalies. The model is trained and evaluated on a custom-labeled drone dataset targeting two key anomaly classes: missing plants, inferred from deviations in planting geometry, and underdeveloped plants, identified through subtle morphological differences. Furthermore, Grad-CAM and transformer attention maps are employed to analyze how models attend to spatial and spectral features, providing interpretability to support decision-making in precision agriculture.

2 Related Work

This section reviews key developments in Salient Object Detection (SOD), from vision-inspired heuristics to deep learning models, with a focus on their adaptation to optical remote sensing imagery (ORSI) and recent applications in agriculture.

2.1 Salient Object Detection (SOD)

Salient Object Detection (SOD) began with biologically inspired models replicating human visual attention, using low-level cues like color, contrast, and orientation. The classic Itti-Koch model computed center-surround differences across image features [10], while later methods like global contrast analysis improved region segmentation by comparing color distributions globally [11]. These early techniques, however, were sensitive to lighting and background complexity, limiting performance in cluttered scenes [12].

Traditional machine learning approaches introduced priors (e.g., center bias, background suppression) and combined features using classifiers like SVMs (Support vector machines) and random forests [13, 9]. Some methods also employed CRFs (Conditional random fields) and optimization frameworks for better spatial coherence [14]. Though efficient, these methods relied on hand-crafted features and lacked robustness in semantically rich environments.

Deep learning methods significantly advanced SOD by enabling hierarchical feature learning. CNN-based models, particularly encoder–decoder and FCN variants, became standard for dense saliency prediction, achieving strong performance in cluttered scenes [15, 13]. However, their application to remote sensing introduces new challenges, such as high-resolution variation, label scarcity, and weak spectral integration.

2.2 ORSI-SOD and Agricultural Applications

Remote sensing imagery presents distinct challenges for saliency modeling. Unlike natural images, RSIs vary in scale, angle, and clutter level. Traditional ORSI-SOD adaptations included sparse Bayesian priors [16], superpixel-based contrast measures [17], and saliency-guided frameworks for building detection [18]. These methods offered some success but were constrained by rigid feature definitions and class-specific tuning.

Deep learning methodology significantly improved ORSI-SOD by leveraging CNNs and transformers to learn global-to-local relationships. Transformer-based saliency detectors better model spatial context, enabling accurate detection across heterogeneous agricultural fields [15]. These models are especially effective when high-resolution inputs and spatial alignment are maintained, though they remain limited by computational cost and generalization issues.

In agriculture, saliency-guided detection has been applied to tasks such as early stress monitoring, weed identification, and yield estimation [19, 20]. Multispectral fusion, particularly RGB-NIR combinations has improved detection of subtle anomalies [21, 22, 23]. Edge-aware refinement and attention mechanisms have further enhanced model interpretability and precision [24]. Lightweight architectures like SggNet support drone-based deployment with limited resources [25]. Despite recent advances, most models continue to struggle with detecting fine-grained issues such as missing or underdeveloped plants. Furthermore, challenges like the limited availability of labeled agricultural datasets and the high computational cost of processing high-resolution imagery remain significant obstacles."

3 Materials and Methods

The proposed approach for detecting regions of interest (ROIs) in agricultural fields utilizes Detection Transformer (DETR) [26], an end-to-end deep learning model that formulates object detection as a direct set prediction problem. DETR simplifies the detection pipeline by directly predicting object bounding boxes and class labels in a single forward pass using a transformer-based encoder-decoder architecture.

DETR-based architectures are particularly well suited for agricultural ROI detection due to their capacity to model complex spatial patterns and capture fine-grained visual cues in high-resolution imagery [27]. Unlike conventional detectors that rely on hand-crafted anchors and region proposals, DETR casts object detection as a sequence prediction task. It segments the image into patches and applies a transformer encoder-decoder to reason holistically across the entire scene [28]. This global attention mechanism allows the model to simultaneously interpret long-range contextual relationships and local visual details. Prior studies demonstrates DETR's strength in detecting nuanced variations in high-resolution remote sensing data [4], highlighting its potential for identifying fine-grained differences, such as underdeveloped crops, as well as coarser anomalies, including planting gaps and structural irregularities. These detections can serve as early indicators of localized field issues, enabling timely and targeted interventions in precision agriculture.

Detecting missing and underdeveloped plants presents two distinct challenges. Missing plants do not differ from the background in appearance and cannot be identified by texture or color. Their only distinguishing characteristic is their unexpected absence from spatially regular planting patterns. Although the grid is not provided directly to the model, DETR's positional encodings enable it to internalize regular spatial structure and detect gaps based on deviations from expected object locations. Underdeveloped plants present a different type of challenge. They are present in the imagery but exhibit only subtle differences from healthy vegetation, such as smaller size, sparser foliage, or a less defined structure. These variations are difficult to detect because they do not stand out strongly in terms of appearance. DETR's global self-attention allows object queries to aggregate information across wide spatial regions, helping the model capture these low-saliency, fine-grained differences by reasoning in context.

By combining positional awareness with global attention, DETR offers a unified approach to detecting both coarse anomalies, like missing plants, and subtle vegetative irregularities associated with underdevelopment.

3.1 Dataset

In order to train and evaluate the proposed detection framework, a custom dataset was constructed using aerial imagery captured by a drone. The dataset includes RGB and NIR images of agricultural fields. This section describes the data collection process, sensor specifications, annotation methodology, and the preparation of ground truth labels used for training and evaluation.

3.1.1 Image Acquisition

The image data used in this study was captured using a DJI Mavic 3 Multispectral drone. This drone is equipped with a high-resolution RGB camera and four narrow-band multispectral sensors (Green, Red, Red Edge, and Near-Infrared), enabling detailed vegetation analysis across both visible and non-visible spectra. This study focuses on the RGB and NIR channels. Images were acquired from agricultural fields during the growing season at altitudes ranging from 12 to 60 meters. The spatial resolution of the images ranged between 4000×3000 and 6000×4000 pixels, providing great detail for identifying plants. The captured fields were characterized by structured plantation layouts with grid-based patterns typical of precision seeding machinery. The scenes included a variety of planting stages, densities, and crop conditions, offering a diverse dataset for robust model training. To ensure consistency across

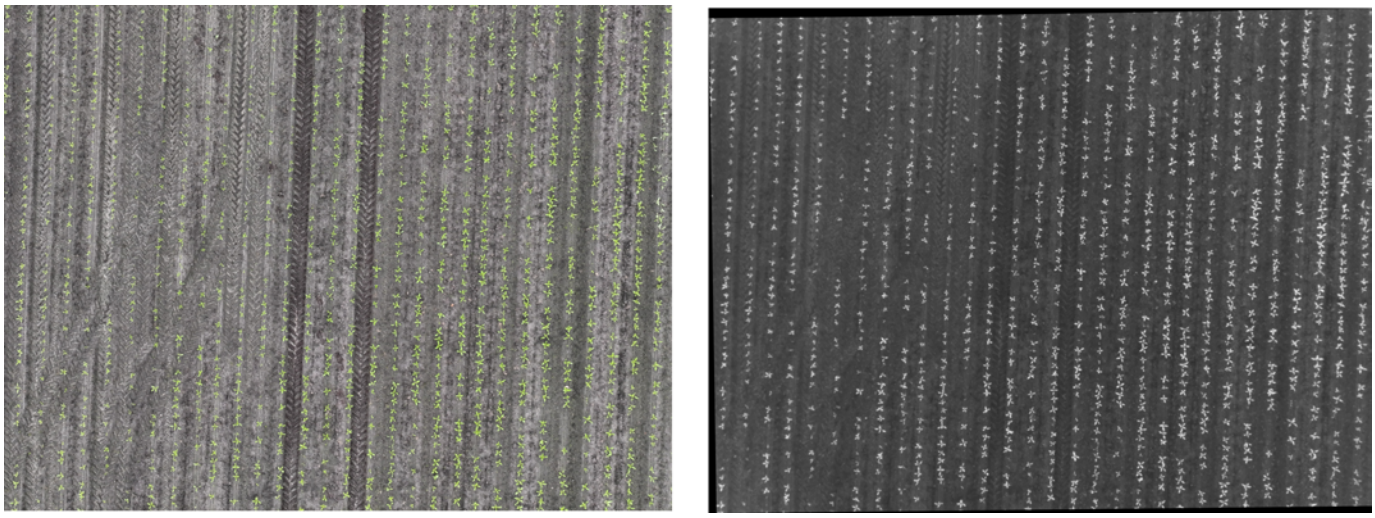


Fig. 1: Aligned and cropped multispectral image pair of an agricultural field. The left image shows the RGB view highlighting visible light features such as plant color and soil texture. The right image shows the corresponding Near-Infrared (NIR) channel, which enhances vegetation contrast and investigated to be useful. Both images are spatially aligned and used jointly in the proposed detection framework.

samples, data from a single growth time zone was used throughout the study.

3.1.2 RGB-NIR Alignment Preprocessing

To ensure effective multispectral fusion, accurate spatial alignment between RGB and Near-Infrared (NIR) imagery is essential. In this study, the alignment procedure is consisted of sensor-specific geometric and radiometric preprocessing, and homography-based alignment refined via feature correspondence.

The first stage involves correcting systematic lens-induced artifacts that affect the physical image formation process. Vignetting correction compensates for radial falloff in illumination by applying a polynomial model parameterized by distance from the calibrated optical center. The correction coefficients and center coordinates are extracted from camera-specific calibration metadata. This step ensures radiometric consistency across the image plane, particularly in the NIR band where peripheral attenuation is more pronounced.

Following vignetting compensation, lens distortion correction is applied to rectify radial and tangential geometric deformations introduced by the camera optics. The mapping from distorted to ideal image coordinates is computed using the intrinsic calibration matrix and distortion coefficients such as focal lengths, principal point offsets, and distortion terms according to the pinhole camera model with radial distortion. This transformation reprojects the observed image onto a geometrically corrected plane, restoring linearity in pixel geometry.

To achieve coarse geometric registration between the RGB and NIR modalities, we apply a projective transformation based on homography decomposition. Two homography matrices are extracted from the drone's onboard calibration metadata: a dewarp matrix mapping the raw sensor image to a designed intermediate plane, and a calibrated transformation mapping the intermediate plane to the RGB reference frame. The composition of these matrices yields a homography that aligns the NIR image to the RGB plane under a planar scene assumption, sufficient for initial spatial consistency in nadir-captured agricultural scenes.

Finally, refined alignment is performed using local feature-based correspondence to compensate for non-linear misalignments arising from inter-band differences in exposure, sensor timing, or residual perspective distortion. Feature points are extracted from the grayscale RGB and pre-aligned NIR images using the Scale-Invariant Feature Transform (SIFT), and matched descriptors are filtered via Lowe's ratio test. A robust estimation of the alignment transformation is then computed using the RANSAC algorithm, allowing for sub-pixel registration accuracy while rejecting spurious matches. The resulting homography further improves alignment fidelity, particularly in fine-structured vegetative regions.

The final output is a pair of co-registered RGB and NIR images, cropped to their common valid region to eliminate peripheral invalid pixels. This ensures that subsequent fusion operations operate on spatially consistent spectral data, which is essential for the integrity of downstream detection and saliency modeling.

3.1.3 Annotations

The annotation process for the dataset was conducted using a semi-automated pipeline, combining algorithmic segmentation with manual inspection and correction to ensure accurate delineation of plant regions in drone imagery. This framework integrated color-based filtering, morphological operations, and watershed segmentation to identify and isolate individual plants across agricultural fields.

Initially, crops were identified based on their spectral characteristics in the HSV (Hue, Saturation, Value) color space. The segmentation was guided by the dominant green hues of young plants, using empirically determined thresholds to isolate vegetation from the background. These thresholds were tuned to maintain generalization across varying illumination conditions while minimizing the inclusion of soil and non-vegetative areas. To refine the binary vegetation masks and reduce noise, a combination of morphological operations was employed. These operations removed small artifacts and closed gaps within crop regions to enhance spatial coherence. The refined masks served as input for plant separation.

To distinguish overlapping or spatially clustered plants, a marker-based watershed segmentation approach was utilized. This technique estimates the most likely boundaries between adjacent vegetative blobs by simulating a topographical flood based on a distance transform. Foreground markers were determined by thresholding the distance map, while uncertain border regions were excluded to reduce over-segmentation. This process produced separated plant instances in complex field layouts where traditional contour-based segmentation would fail.

For each isolated plant region, a bounding box was defined that tightly encapsulated the area. To compensate for minor undersegmentation and ensure full plant coverage, each bounding box was systematically expanded by a fixed percentage in both dimensions. This approach preserved alignment with the plant's true extent without extending beyond image boundaries. Bounding boxes with areas below a defined threshold were discarded as noise or

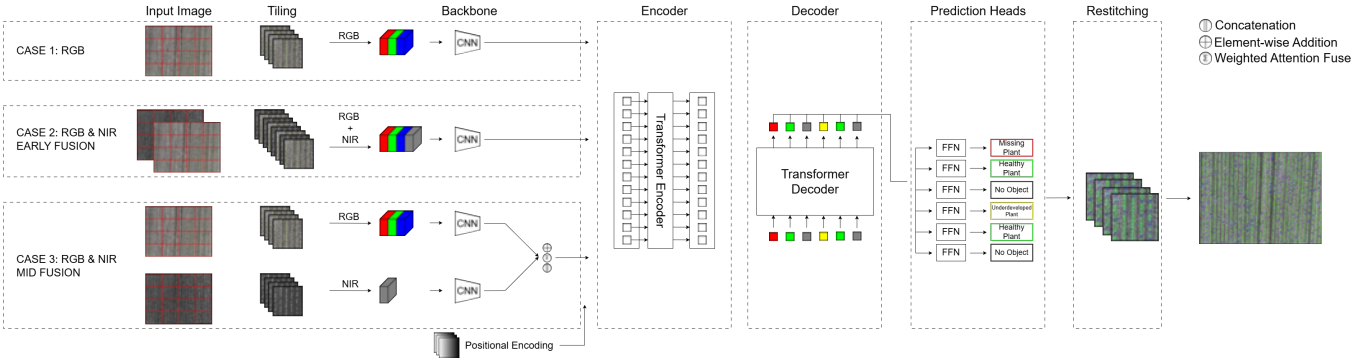


Fig. 2: Overview of the proposed customized DETR-based architecture for multispectral agricultural field analysis. Input RGB and NIR images are tiled and preprocessed into various combinations (RGB, NIR, and RGB+NIR). Each tile is passed through a CNN backbone, followed by fusion before being processed by a Transformer Encoder. The Transformer Decoder generates object queries used by parallel feed-forward networks (FFNs) to classify detected regions into categories: Healthy Plant, Underdeveloped Plant, Missing Plant, or No Object. Finally, the predictions are re-stitched to reconstruct the full field-level detection map.

non-representative of actual crops.

Beyond labeling visible plant regions, missing plant zones were annotated manually using spatial priors derived from plantation geometry. In structured agricultural fields, precision seeders deposit plants at fixed radial or linear intervals, forming regular circular or grid-like patterns. These patterns were used to infer expected plant locations. Annotators overlaid a virtual planting grid derived from field metadata and marked positions where a plant was absent. This grid was not used as an input to the model but served as an external geometric reference to support consistent and objective annotation of missing entities.

To ensure reproducibility, all parameter values used in segmentation and post-processing steps are detailed in Appendix A.

3.2 Detection Strategy

The detection pipeline is built on the Detection Transformer (DETR), an end-to-end object detection architecture that combines convolutional feature extraction with transformer-based global reasoning. A ResNet backbone, such as ResNet-50 or ResNet-101, extracts high-level feature maps from the input image. These features are enriched with positional encodings to retain spatial structure and passed to a transformer encoder, which models long-range dependencies through multi-head self-attention. A fixed set of learnable object queries in the decoder then attends to the encoded features to predict object classes and bounding boxes.

While DETR's core architecture remains unchanged, three key adaptations are introduced to suit the demands of drone-based agricultural monitoring: (1) spatially preserving image tiling to handle high-resolution inputs, (2) multispectral feature extraction via early and mid-level fusion of RGB and NIR channels, and (3) a custom post-processing stage for merging tile-based predictions using confidence filtering and non-maximum suppression. These modifications ensure that fine-grained plant-level anomalies can be detected effectively across large field-scale scenes.

3.2.1 Image Tiling

To efficiently process high-resolution drone imagery while preserving critical spatial detail, each input image is divided into overlapping tiles of size 800×800 pixels with a stride of 600 pixels. This sliding window approach balances computational feasibility with contextual completeness, ensuring that objects near tile borders are captured in multiple contexts. It avoids the resolution loss associated with global downscaling, which would hinder the detection of small or underdeveloped plants.

Given an image of size $H \times W$, the number of tiles is computed as:

$$N_H = \left\lfloor \frac{H-800}{600} \right\rfloor + 1, \quad N_W = \left\lfloor \frac{W-800}{600} \right\rfloor + 1. \tag{1}$$

For images not evenly divisible by the stride, the remaining edge regions are zero-padded to ensure full coverage. This tiling strategy enables GPU-efficient training and inference while maintaining spatial fidelity for precise anomaly detection.

3.2.2 Feature Extraction with Convolutional Backbone

Visual features are extracted from each image tile using a ResNet backbone [29], pretrained on ImageNet and modified for multispectral input. In the baseline configuration, RGB tiles are processed by a standard ResNet-50 or ResNet-101, producing semantically rich feature maps with reduced spatial resolution.

To incorporate Near-Infrared (NIR) data, two fusion strategies are implemented. In early fusion, the RGB and NIR channels are concatenated into a 4-channel input, and the first convolutional layer is adjusted to accept this input, initializing NIR weights with either averaged RGB filters or random initialization. This enables joint learning from all channels from the outset.

In mid-level fusion, RGB and NIR tiles are processed by parallel ResNet backbones. Their feature maps are then merged at an intermediate stage using either additive fusion, channel-wise concatenation, or attention-based weighting. This preserves modality-specific features while enabling joint reasoning prior to transformer encoding.

The resulting feature maps, whether from single- or dual-stream configurations, are flattened into embeddings and passed to the transformer encoder. Spatial alignment between RGB and NIR inputs is ensured through preprocessing to maintain consistency during fusion.

3.2.3 Transformer Reasoning and Prediction

The fused feature maps are passed to a transformer encoder, which applies multi-head self-attention across all spatial positions to model long-range dependencies and contextual relationships in the image. The decoder operates over a fixed set of learnable object queries that iteratively interact with the encoded features through cross-attention, producing object-specific embeddings. Each embedding is then decoded into a class label and bounding box via feed-forward networks. This attention-based formulation allows the model to reason globally and detect both explicit plant regions and implicit anomalies such as missing instances.

3.2.4 Inference and Post-Processing

At inference time, each image tile is independently passed through the trained DETR model, which produces a fixed set of N_q object queries per tile:

$$\hat{y} = \text{DETR}(I; \theta) \tag{2}$$

where θ denotes the learned model parameters. Each query predicts a class label and a bounding box. Since the number of queries is fixed, some predictions may correspond to background or low-confidence outputs.

To reconstruct full-image detections, tile-local predictions are filtered by a confidence threshold τ and projected to global coordinates using the tile's top-left offset:

$$\mathscr{P} = \{(b_i, c_i, s_i)\}, \quad i = 1, ..., N_d \tag{3}$$

where:

- $b_i \in \mathbb{R}^4$ is the predicted bounding box in global coordinates,
- $c_i \in \mathbb{Z}^+$ is the predicted class label,
- $s_i \in [0,1]$ is the confidence score,
- N_d is the number of detections with $s_i \geq \tau$.

To resolve overlapping detections from adjacent tiles, a custom non-maximum suppression (NMS) scheme is applied. It evaluates both Intersection over Union (IoU) and containment ratio between bounding boxes, suppressing redundant detections based on class-specific priority rules. This step ensures a spatially coherent and globally consistent set of predictions.

3.2.5 Evaluation Metrics

Model performance is evaluated using four standard object detection metrics: Precision, Recall, F1 Score, and mean Average Precision (mAP). These metrics are computed over varying confidence thresholds and Intersection over Union (IoU) levels to assess both classification accuracy and spatial localization.

Precision and **Recall** are defined as:

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN} \tag{4}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The **F1 Score** combines both measures:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

The **mean Average Precision (mAP)** summarizes detection quality across classes and recall levels:

$$\text{mAP} = \frac{1}{C} \sum_{i=1}^{C} \text{AP}_i \tag{6}$$

where C is the number of classes and AP_i is the average precision for class i. Evaluations are performed at IoU thresholds of 0.5, 0.75, and the averaged range from 0.5 to 0.95.

4 Experiments and Results

This section presents the experimental setup, evaluation protocols, and performance results for the proposed detection framework. The experiments are designed to assess the effectiveness of various architectural configurations.

4.1 Experimental Setup

Three primary configurations of core DETR architecture is explored: (i) a baseline RGB-only pipeline, (ii) an early fusion model integrating Near-Infrared (NIR) data as an extra input channel, and (iii) a mid-level fusion model employing dual-stream ResNet backbones for RGB and NIR.

Models are trained for 600 epochs, with evaluations performed every 5 epochs. To mitigate overfitting, early stopping is employed with a patience of 10 evaluation intervals. Optimization is carried out using the AdamW algorithm. Parameters outside the backbone are updated with a learning rate of 1×10^{-4}, while backbone layers use a lower rate of 1×10^{-5}. In mid-fusion configurations, the NIR backbone is initialized independently and trained with a learning rate of 1×10^{-4}, while the RGB backbone retains the lower rate of 1×10^{-5}, reflecting its pretraining on ImageNet. A batch size of 4 is used. A learning rate scheduler based on validation performance is employed, specifically the ReduceLROnPlateau strategy, which monitors evaluation loss and adaptively decreases the learning rate by a factor of 0.1 when improvement stagnates, thereby promoting continued convergence.

Model performance is evaluated using mean Average Precision (mAP), precision, recall, and F1-score. These metrics are computed separately for the plant and missing-plant categories at multiple IoU thresholds, including 0.5, 0.75, and the averaged range from 0.5 to 0.95 with a step size of 0.05.

4.2 Backbone and Checkpoint Variants

To assess the impact of backbone depth and dilation, four DETR variants were benchmarked at IoU = 0.50 using RGB input. As shown in Table 1, DETR-R50 achieved the best balance of precision, recall, and F1 score. While DETR-DC5-R101 slightly outperformed others in mAP, the margin was negligible.

These results indicate that increasing backbone complexity does not lead to meaningful improvements when high-resolution tiling is applied. The spatial detail preserved through tiling enables DETR-R50 to perform competitively, making it the most efficient choice for this task.

Table 1: Detection performance of DETR variants with different ResNet backbones using RGB input at IoU = 0.50. All models are initialized with ImageNet-pretrained weights and fine-tuned on the agricultural dataset. The best result for each metric is highlighted in bold.

Model	Precision	Recall	F1 Score	mAP
DETR-R50	**0.8314**	**0.8145**	**0.8227**	0.5067
DETR-DC5-R50	0.8144	0.8086	0.8112	0.4989
DETR-R101	0.8256	0.8012	0.8130	0.5048
DETR-DC5-R101	0.8233	0.8118	0.8172	**0.5068**

4.3 Positional Encoding Strategies

This experiment compares fixed sinusoidal encodings with learned positional embeddings to evaluate spatial reasoning in structured agricultural fields. As shown in Table 2, sinusoidal encodings yield substantially higher detection performance at IoU = 0.50, with over 3 times improvement in precision and nearly 20 times in mAP.

The results highlight that learned encodings fail to capture spatial regularities, likely due to limited data and the uniform grid structure of crop layouts. In contrast, fixed encodings provide a strong spatial prior that significantly enhances detection reliability, underscoring their importance in transformer-based models for this domain.

Table 2: Comparison of sine-based and learned positional encodings in DETR using RGB input at IoU = 0.50. All models are trained on the same dataset with identical settings, differing only in the type of positional embedding used.

Model	Precision	Recall	F1 Score	mAP
Sinusodial PE	**0.8314**	**0.8145**	**0.8227**	**0.5067**
Learned PE	0.2184	0.1863	0.1904	0.0264

4.4 Single-Channel Spectral Band Training

This experiment isolates the contribution of each visible band by training DETR on single-channel inputs replicated across all three channels. As shown in Table 3, the green channel alone (DETR-G) achieves F1 and mAP values closely approaching the RGB baseline.

This indicates that it carries the most relevant information for plant detection among the visible bands.

The red channel performs moderately well, while the blue channel achieves lower detection performance, suggesting that it contributes less to identifying vegetative structures in this context. Overall, the results demonstrate that the green band can serve as an effective alternative when full-spectrum RGB input is not available, although combining all three channels still offers a slight advantage.

Table 3: Detection performance under different spectral band configurations using DETR at IoU = 0.50.

Model	Precision	Recall	F1 Score	mAP
RGB (Baseline)	**0.831**4	**0.8145**	**0.8227**	**0.5067**
Red (DETR-R)	0.8066	0.7968	0.8013	0.4589
Green (DETR-G)	0.8255	0.8093	0.8169	0.5005
Blue (DETR-B)	0.8063	0.7919	0.7986	0.4660

4.5 NIR-Only Detection

To assess the standalone utility of Near-Infrared input, DETR was trained using only the NIR band replicated across channels. As shown in Table 4, the NIR-only model achieves 69% of the RGB baseline in mAP and underperforms notably in F1 score.

These results confirm that while NIR is proposed to capture vegetation-background contrast, it lacks the performance required for fine-grained anomaly detection. NIR alone is insufficient for the task and is better suited as a complementary modality in multispectral fusion.

Table 4: Detection performance comparison between DETR models trained on NIR-only input and the RGB baseline at IoU = 0.50.

Model	Precision	Recall	F1 Score	mAP
RGB (Baseline)	**0.831**4	**0.8145**	**0.8227**	**0.5067**
NIR Only	0.7514	0.7080	0.7287	0.3484

4.6 Early Fusion Configurations

This experiment investigates input-level fusion of RGB and NIR by comparing full 4-channel fusion with reduced 2-channel combinations. As shown in Table 5, all early fusion variants underperform compared to the RGB baseline.

Full RGB+NIR models yield a consistent drop in F1 score and mAP, suggesting that direct concatenation of NIR introduces poorly aligned or redundant features. Even the best 4-channel setup, EF (NIR First), does not recover this loss.

In contrast, simpler 2-channel fusions such as G+NIR perform better than their 4-channel counterparts, with G+NIR reaching the highest early fusion score. Still, none surpass RGB-only performance. These results suggest that early fusion of NIR is not inherently beneficial and must be guided by selective spectral design or more structured integration mechanisms.

Table 5: Performance of early fusion strategies at IoU = 0.50, including full RGB+NIR and single RGB band + NIR configurations.

Model	Precision	Recall	F1 Score	mAP
RGB (Baseline)	**0.8314**	**0.8145**	**0.8227**	**0.5067**
EF (NIR Last)	0.7959	0.7765	0.7857	0.4674
EF (NIR First)	0.8023	0.7767	0.7889	0.4638
EF (R+NIR)	0.7254	0.6732	0.6971	0.3559
EF (G+NIR)	0.7535	0.7201	0.7356	0.3925
EF (B+NIR)	0.7276	0.6734	0.6982	0.3553
EF (RRR+NIR)	0.6533	0.5797	0.6114	0.2699
EF (GGG+NIR)	0.6386	0.5760	0.6031	0.2888
EF (BBB+NIR)	0.4977	0.4192	0.4494	0.1791

4.7 Mid-Level Fusion

This experiment explores structured fusion by processing RGB and NIR through separate backbones before merging features. As shown in Table 6, element-wise addition (*MF (Add)*) achieves the highest mAP, slightly exceeding the RGB baseline, but does not improve the F1 score.

Attention-weighted fusion yields comparable results but introduces additional complexity without clear benefit. Channel-wise concatenation performs on par with the RGB baseline, indicating that naive feature stacking does not degrade performance but also does not provide a measurable advantage.

Fusion with individual RGB bands reinforces the dominance of the green channel, yet no mid-level fusion configuration surpasses the RGB baseline meaningfully except in mAP. While structured fusion through mid-level integration offers better results than early fusion, the practical benefits remain limited in this setting.

Table 6: Detection performance at IoU = 0.50 for mid-level fusion strategies. Fusion methods include element-wise addition, channel-wise concatenation, and attention-weighted fusion. Additional rows show fusion of NIR with individual RGB bands.

Model	Precision	Recall	F1 Score	mAP
RGB (Baseline)	0.8314	**0.8145**	0.8227	0.5067
MF (Add)	**0.8405**	0.8087	**0.8241**	**0.5168**
MF (Attention)	0.8338	0.8055	0.8192	0.4872
MF (Concat)	0.8378	0.8093	0.8231	0.5160
MF (RRR+NIR)	0.8164	0.8015	0.8086	0.4617
MF (GGG+NIR)	0.8167	0.8036	0.8098	0.4903
MF (BBB+NIR)	0.8081	0.7922	0.7998	0.4535

4.8 Resolution Degradation Study

This experiment investigates the impact of global image resolution on detection accuracy. Original drone images were uniformly downsampled using scale factors from 0.9 to 0.5, with tiling applied post-resizing to reflect true spatial loss.

As shown in Figure 3, precision remains relatively robust across scales, but recall, F1 score, and mAP decline sharply below a resize factor of 0.8. The steepest drop occurs between 0.7 and 0.6, indicating a critical resolution threshold for reliable anomaly detection.

These results suggest that reducing input resolution too aggressively impairs the model's ability to localize and classify fine-grained features. To maintain detection quality, a scale factor of 0.8 or higher is recommended, offering practical guidance for flight altitude and imaging strategy in drone-based agricultural monitoring.

4.9 Model Explainability and Attention Analysis

To better understand the decision-making process of the detection models, we apply interpretability techniques focused on saliency and

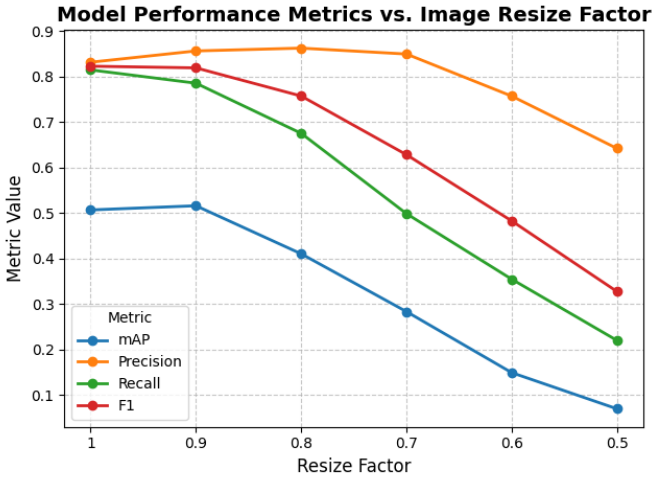


Fig. 3: Effect of global image downscaling on mAP, precision, recall, and F1 score at IoU = 0.50. Sharp drops are observed below the 0.8 scaling factor, especially in recall and F1.

attention behavior. This section analyzes how different model configurations attend to image regions.

4.9.1 Grad-CAM Visualization

To analyze how models localize fine-grained cues, we apply Gradient-weighted Class Activation Mapping (Grad-CAM) to the convolutional backbones, focusing on the "underdeveloped plant" class. These instances are visually subtle, requiring sensitivity to reduced size, density, or irregular spacing.

We visualize Grad-CAM heatmaps for five model types: RGB-only, NIR-only, Green-only, Early Fusion, and Mid-Level Fusion. For single-stream models, Grad-CAM is applied to the final convolutional block, while for mid-fusion, RGB and NIR branches are visualized separately. This allows direct comparison of spatial focus and saliency behavior across architectures and input modalities.

The Grad-CAM visualizations in Fig. 4 highlight how different input configurations influence spatial focus during underdeveloped plant detection. RGB and green-only models show the most localized and structured activations, concentrating on plant contours and intra-row patterns. These areas correspond to subtle morphological cues, indicating that both inputs effectively support fine-grained detection. The RGB model exhibits slightly sharper focus, likely due to its richer spatial context.

In contrast, NIR-only input results in diffuse and poorly defined saliency, with activations spread across the scene. This lack of focus aligns with its lower detection performance and suggests limited spatial utility when used alone. Early fusion introduces inconsistent saliency, with scattered activation across vegetation and background, reflecting unstructured signal mixing. Mid-level fusion improves focus in the RGB stream but shows minimal contribution from the NIR branch, suggesting the model underutilizes NIR features when RGB dominates training.

Overall, the results demonstrate that visible spectrum inputs are more effective for identifying underdeveloped crops and that NIR requires more targeted integration to be beneficial.

4.9.2 Transformer Attention Maps

To understand how the transformer-based detector models spatial relationships, we analyze encoder and decoder attention maps for the underdeveloped plant class. While Grad-CAM highlights influential regions in convolutional features, attention maps reveal how predictions are formed through spatial interactions.

Figure 5 shows decoder attention (object queries attending to key regions) and encoder self-attention (how spatial tokens interact

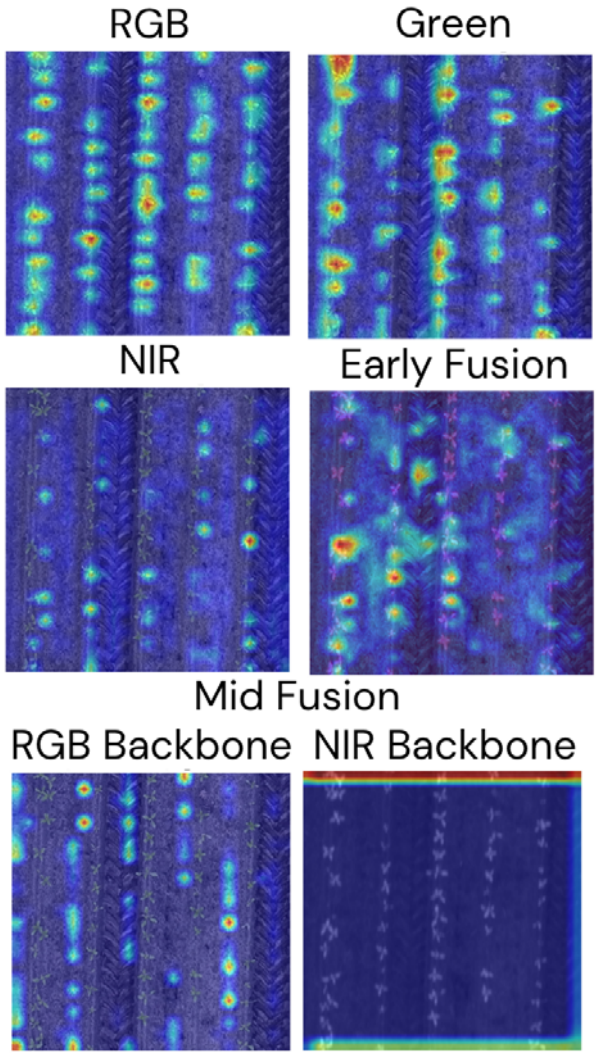


Fig. 4: Class-specific Grad-CAM visualizations for the "underdeveloped plant" category across different model configurations. Heatmaps indicate the spatial focus of each model when identifying subtle signs of plant underdevelopment.

globally). This dual view illustrates how models balance local focus and contextual reasoning in dense agricultural layouts.

Inputs based on the visible spectrum, particularly RGB and green-only configurations, yield the most structured and localized attention. Decoder queries in these models consistently align with underdeveloped plant regions, while encoder self-attention follows row patterns and supports context-aware inference. This reflects strong spatial understanding and aligns with the high detection performance of these models.

In contrast, the NIR-only configuration results in weak and dispersed attention. Decoder activations frequently miss target regions, and encoder attention lacks clear structure, indicating that NIR alone does not offer sufficient spatial cues for fine-grained detection. Fusion strategies show varying effectiveness. Early fusion leads to inconsistent attention behavior, with scattered focus across vegetated and non-vegetated areas. This suggests that simple channel stacking introduces noise and reduces spatial specificity. Mid-level fusion shows more structured decoder and encoder attention, similar to RGB-only models, suggesting that separate processing of modalities helps retain spatial focus.

Overall, the attention patterns emphasize the strength of visible-spectrum inputs in guiding transformer attention and spatial

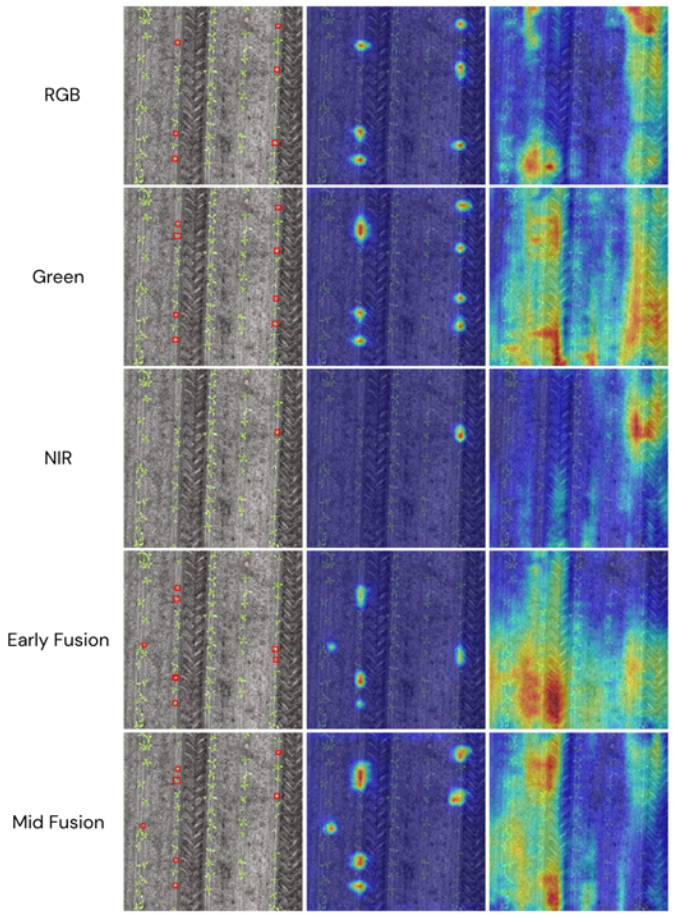


Fig. 5: Transformer attention visualizations from the final encoder and decoder layers across different input configurations.

reasoning. While fusion strategies may offer some benefits in other settings, they fall behind in this comparison. Their effectiveness depends heavily on preserving spatial consistency across modalities, which remains challenging. Attention visualizations thus offer a valuable window into how different input configurations shape the model's internal representation of spatial dependencies.

5 Discussion, Conclusion & Future Work

This section reflects on the key findings of the study, placing them in the context of the broader goals of precision agriculture and multispectral object detection. We discuss the implications of the experimental results, including the performance of various spectral inputs and fusion strategies, and how these findings inform model interpretability and practical deployment. The section concludes with a summary of contributions and outlines potential directions for future research to further enhance detection performance and robustness.

5.1 Discussion

This study demonstrates that transformer-based object detection, particularly DETR, is effective for identifying regions of interest in structured crop fields. These include both healthy plants and irregularities such as underdeveloped or missing vegetation. RGB-based configurations showed consistently strong performance across IoU thresholds. The green channel alone approached full RGB accuracy, highlighting its relevance for visual distinction between plant types. Incorporating Near-Infrared (NIR) through mid-level fusion led to slight improvements in mAP and precision, but the added preprocessing, image handling, and architectural complexity may not be justified. Early fusion strategies performed worse than the baseline, indicating that naive input-level integration of NIR is not effective.

Grad-CAM analysis showed that RGB and green-only models focused sharply on plant boundaries and row structure, while early fusion displayed scattered activations that failed to localize vegetation consistently. The NIR backbone in mid-level fusion showed weak and diffuse responses, suggesting underutilization of the spectral modality.

Slightly reducing image resolution by 10 to 20% led to a modest precision gain. This may be due to more plant instances fitting within each tile, giving the model richer spatial context during training. Since spatial detail was largely preserved, performance remained stable. The reduced input size also lowered computational load, making training more efficient. However, performance dropped sharply when resolution fell below approximately 80%, indicating a lower bound for reliable detection.

There are also several limitations. Multispectral preprocessing, particularly the alignment of RGB and NIR images, significantly increases inference time. This step introduces a tenfold increase in inference time compared to RGB-only models, rendering NIR-based configurations impractical for real-time or on-device use. In contrast, RGB-only inference is nearly instantaneous. Dataset size was also limited by the manual annotation process. Each image contains more than 4,000 plants. Annotating missing and underdeveloped plants required full manual labeling, and healthy labels needed manual validation after segmentation. This constrained data diversity and scale. Additionally, the explored fusion strategies did not fully utilize the NIR input, suggesting the need for more adaptive integration methods.

While the results offer useful insights into model behavior, it is important to note that performance differences have not been tested for statistical significance. Ideally, each model variant would be run multiple times to estimate variance and confidence intervals. This was not feasible due to the high computational cost of training and evaluation. As a result, small metric differences, especially under one or two percentage points, should be interpreted with caution. Consistent trends, such as those seen in resolution experiments or spectral ablations, are more likely to reflect meaningful effects.

5.2 Conclusion

This study demonstrates that transformer-based object detection, specifically DETR, can be successfully adapted to identify both missing and underdeveloped plants in structured agricultural fields. By combining high-resolution tiling with attention-based interpretability, the approach delivers reliable detection of subtle and large-scale anomalies, showing strong suitability for analyzing drone-captured imagery in precision agriculture settings.

The investigation into spectral input configurations revealed that RGB-based models consistently outperformed alternatives, with the green channel alone capturing most of the discriminative visual features. While Near-Infrared data was explored through early and mid-level fusion, it offered only marginal benefit and introduced significant overhead in preprocessing and model complexity. These findings suggest that visible spectrum cues are sufficient for morphology-based plant anomaly detection in this context and that additional modalities require more adaptive integration strategies to be effective.

Image resolution was also found to be a critical factor. Slight reductions maintained or even improved performance, likely due to increased spatial context within tiles, but detection quality dropped sharply below a certain threshold. This provides actionable guidance for flight planning and data acquisition in field deployments, ensuring that spatial detail is preserved without excessive computational cost.

Beyond performance metrics, the framework offers interpretable outputs that strengthen its value in real-world use. Attention visualizations revealed how the model focuses on planting geometry and vegetation structure when making predictions. These saliency cues can help agronomists understand not only what the model detects but also why, potentially exposing issues like compaction

zones, irrigation inconsistencies, or emerging stress that are not explicitly labeled.

Real-world deployment of this system must also consider broader implications. The use of transparent, interpretable AI outputs supports ethical deployment, particularly when decisions such as replanting or treatment are guided by model predictions. At the legal level, drone-based monitoring involves the collection of high-resolution data over private land, raising questions around consent, ownership, and data governance. From a societal perspective, there is a risk of reinforcing inequality if these technologies remain accessible only to large-scale operations. Supporting open implementations and efficient models is key to extending their benefits to smallholder farms and public institutions.

Together, the findings contribute a robust, interpretable, and field-ready detection approach that aligns with the goals of scalable, sustainable, and responsible precision agriculture.

5.3 Future Work

Future research should begin with a full statistical validation of the results presented in this study. Each model configuration should be trained and evaluated multiple times to report confidence intervals and variance to confirm whether observed trends are statistically significant. Such validation should be completed before extending the architecture or building new research directions on these findings.

Another direction is to expand the anomaly categories addressed by the framework. Detecting early-stage disease symptoms, pest damage, and nutrient deficiencies would increase the relevance of the system for broader crop monitoring tasks.

Improving annotation efficiency remains a key priority. Using the current model to assist in labeling can help scale the dataset and allow for more diverse research to advance the state of the art in the field.

The limited contribution of NIR observed in this study also points to the need for more advanced fusion mechanisms. Further research should investigate adaptive fusion strategies, such as cross-modality attention mechanisms and multi-scale calibration modules [30], as well as gated fusion architectures with modality-specific Transformer encoders [31]. Such approaches may better exploit the complementary information in NIR and other spectral bands when relevant.

For real-time deployment, RGB models offer a feasible path due to minimal preprocessing. In contrast, multispectral setups require improved alignment strategies, as current methods introduce significant delay. Future work should focus on faster, learnable alignment to make multispectral use practical in time-constrained settings.

Selective band replacement is another promising direction. Instead of combining NIR with RGB, future work could investigate replacing one RGB band with NIR to form configurations like NIR+R+G. This may retain critical visual features while introducing complementary spectral information.

Acknowledgements

- This project is financially supported by the European Union within the JTF-call project DigiAgro3



References

[1] Food and Agriculture Organization of the United Nations. The impact of disasters on agriculture and food security 2023: Avoiding and reducing losses through investment in resilience, 2023.

[2] Marco Cantillo, Andrea Cattaneo, Abram Bicksler, Adriana Ignaciuk, Giorgio Grussu, Yuka Makino, Dario Lucantoni, Anne Mottet, Beate Scherf, Antonio Scognamillo, and Fenton Beed. *The State of Food and Agriculture 2021 agrifood systems resilient to shocks and stresses*. 11 2021.

[3] Fluturim Saliu, Muhammad Luqman, and Hasan AlKhaza'leh. A review on the impact of sustainable agriculture practices on crop yields and soil health. *International Journal of Information Technology*, 2:1–13, 05 2023.

[4] Jiahui Xu, Yuanpei Cui, Shuhan Zhang, and Meiping Zhang. The evolution of precision agriculture and food safety: a bibliometric study. *Frontiers in Sustainable Food Systems*, 8, 2024.

[5] Sewnet Getahun, Habtamu Kefale, and Yohannes Gelaye. Application of precision agriculture technologies for sustainable crop production and environmental sustainability: A systematic review. *The Scientific World Journal*, 2024:2126734, 2024.

[6] Hao Han, Zhiwei Liu, Jianfeng Li, and Zhiwei Zeng. Challenges in remote sensing-based climate and crop monitoring: Navigating the complexities using ai. *Journal of Cloud Computing*, 13(1):83, 2024.

[7] Emmanuel Omia, Hyungjin Bae, Eunsung Park, Moon Sung Kim, Insuck Baek, Isa Kabenge, and Byoung-Kwan Cho. Remote sensing in field crop monitoring: A comprehensive review of sensor systems, data analyses and recent advances. *Remote Sensing*, 15(2), 2023.

[8] Minhui Li, Redmond R. Shamshiri, Cornelia Weltzien, and Michael Schirrmann. Crop monitoring using sentinel-2 and uav multispectral imagery: A comparison case study in northeastern germany. *Remote Sensing*, 14(17), 2022.

[9] Longxuan Yu, Xiaofei Zhou, Lingbo Wang, and Jiyong Zhang. Boundary-aware salient object detection in optical remote-sensing images. *Electronics*, 11(24), 2022.

[10] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[11] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu. Global contrast based salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 409–416, 2011.

[12] Ashish Kumar Gupta, Ayan Seal, Mukesh Prasad, and Pritee Khanna. Salient object detection techniques in computer vision—a survey. *Entropy*, 22(10):1174, October 2020.

[13] Xiuwenxin Wang, Siyue Yu, Eng Gee Lim, and M L Dennis Wong. Salient object detection: A mini review. In *Frontiers in Signal Processing*, volume 4. Frontiers Media SA, May 2024.

[14] Xiang Wang, Huimin Ma, and Xiaozhi Chen. Geodesic weighted bayesian model for saliency optimization. *Pattern Recognition Letters*, 75:1–8, 2016.

[15] Gongyang Li, Zhen Bai, Zhi Liu, Xinpeng Zhang, and Haibin Ling. Salient object detection in optical remote sensing images driven by transformer. *arXiv [cs.CV]*, 2023.

[16] Danpei Zhao, Jiajia Wang, Jun Shi, and Zhiguo Jiang. Sparsity-guided saliency detection for remote sensing images. *Journal of Applied Remote Sensing*, 9(1):095055, 2015.

[17] Long Ma, Bin Du, He Chen, and Nouman Q. Soomro. Region-of-interest detection via superpixel-to-pixel saliency analysis for remote sensing image. *IEEE Geoscience and Remote Sensing Letters*, 13(12):1752–1756, 2016.

[18] Er Li, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Building extraction from remotely sensed images by integrating saliency cue. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(3):906–919, 2017.

[19] Jingzong Zhang, Shijie Cong, Gen Zhang, Yongjun Ma, Yi Zhang, and Jianping Huang. Detecting pest-infested forest damage through multispectral satellite imagery and improved unet++. *Sensors*, 22(19):7440, Sep 2022.

[20] Anum Naveed, Wasif Muhammad, Muhammad Irshad, Javaid Aslam, Sajjad Manzoor, Tasleem Kausar, and Yun Lu. Saliency-based semantic weeds detection and classification using uav multispectral imaging. *IEEE Access*, PP:1–1, 01 2023.

[21] Zuojun Zheng, Jianghao Yuan, Wei Yao, Paul Kwan, Hongxun Yao, Qingzhi Liu, and Leifeng Guo. Fusion of uav-acquired visible images and multispectral data by applying machine-learning methods in crop classification. *Agronomy*, 14(11), 2024.

[22] Nuno Cunha, Tiago Barros, Mário Reis, Tiago Marta, Cristiano Premebida, and Urbano J. Nunes. Multispectral image segmentation in agriculture: A comprehensive study on fusion approaches, 2023.

[23] Ayyappa Reddy Allu and Shashi Mesapam. Impact of remote sensing data fusion on agriculture applications: A review. *European Journal of Agronomy*, 164:127478, 2025.

[24] Hongli Li, Xuhui Chen, Wei Yang, Jian Huang, Kaimin Sun, Ying Wang, Andong Huang, and Liye Mei. Global semantic-sense aggregation network for salient object detection in remote sensing images. *Entropy*, 26(6), 2024.

[25] Jie Liu, Jinpeng He, Huaixin Chen, Ruoyu Yang, and Ying Huang. A lightweight semantic- and graph-guided network for advanced optical remote sensing image salient object detection. *Remote Sensing*, 17(5), 2025.

[26] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020.

[27] Ranjan Sapkota, Rahul Harsha Cheppally, Ajay Sharda, and Manoj Karkee. Rf-detr object detection vs yolov12: A study of transformer-based and cnn-based architectures for single-class and multi-class greenfruit detection in complex orchard environments under label ambiguity, April 17 2025. arXiv:2504.13099v1.

[28] Gracile Astlin Pereira and Muhammad Hussain. A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships, August 27 2024. arXiv:2408.15178v1.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[30] Xijian Fan, Chunlei Ge, Xubing Yang, and Weice Wang. Cross-modal feature fusion for field weed mapping using rgb and near-infrared imagery. *Agriculture*, 14(12), 2024.

[31] Zeynep Galymzhankyzy and Eric Martinson. Lightweight multispectral crop-weed segmentation for precision agriculture, 05 2025.

A HSV Segmentation and Annotation Parameters

This appendix outlines the full set of parameters used in the semi-automated annotation pipeline for detecting and segmenting plant regions from drone imagery. These parameters were selected to balance segmentation accuracy and robustness under varying field and lighting conditions.

1. HSV Color Filtering

To isolate vegetation based on hue and chromatic properties:

- **Lower HSV Bound:** `[30, 20, 40]`
- **Upper HSV Bound:** `[90, 255, 255]`

These thresholds target dominant green hues typical of young vegetation while avoiding soil, shadow, or dried foliage.

2. Morphological Processing

Morphological operations are applied to clean the binary vegetation masks:

- **Structuring Kernel Size:** 3×3
- **Opening (Noise Removal):** `cv2.MORPH_OPEN`
- **Closing (Gap Filling):** `cv2.MORPH_CLOSE`

These steps remove small speckles and bridge broken segments within detected plant regions.

3. Distance Transform and Marker Extraction

Foreground markers for watershed segmentation are derived using distance-based foreground estimation:

- **Distance Transform Type:** Euclidean (`cv2.DIST_L2`)
- **Distance Transform Kernel Size:** 5
- **Foreground Thresholding:** $0.3 \times \max(\texttt{dist_transform})$
- **Marker Generation:** `cv2.connectedComponents()`
- **Unknown Region Label:** Pixels set to 0 in the marker matrix

These settings ensure robust separation of spatially clustered plants using a topographic flooding analogy.

4. Watershed Segmentation

The watershed algorithm is applied to segment individual plant instances from the foreground mask:

- **Watershed Algorithm:** `cv2.watershed()`
- **Input Image:** 3-channel RGB image used for watershed background coloring
- **Contour Drawing Color (for visualization):** `(255, 0, 0)`

5. Bounding Box Extraction and Post-processing

To create bounding boxes from segmented plant masks:

- **Minimum Bounding Box Area:** 50 pixels (boxes below this are discarded as noise)
- **Bounding Box Expansion Factor:** 10% (applied to both width and height symmetrically)
- **Clipping:** Bounding boxes are clipped to remain within image boundaries

All parameter values were fixed for all scenes and used consistently to ensure reproducibility of segmentation and annotation across the dataset.

B Detailed Evaluation Metrics

This appendix presents detailed evaluation metrics for all tested DETR-based model configurations. The results provide a comprehensive view of detection performance across different Intersection over Union (IoU) thresholds and for each target class. In addition to grouped metrics (overall mAP, precision, recall, and F1), class-wise performance is reported to illustrate how well each model handles specific anomaly categories: Healthy Plant, Underdeveloped Plant, and Missing Plant. These extended results complement the main text and enable deeper analysis of model behavior under varying configurations and input conditions.

Table 7: Grouped evaluation metrics for DETR-based models across different IoU thresholds. The table reports overall mAP, precision, recall, and F1 score at IoU = 0.50, IoU = 0.75, and averaged IoU = 0.50:0.95. Best results are highlighted in bold. These metrics reflect the overall detection quality for each model variant across all target classes.

Model	**IoU:0.5**				**IoU:0.75**				**IoU:(0.5:0.95)**			
	mAP	P	R	F1	mAP	P	R	F1	mAP	P	R	F1
RGB (R50)	0.5067	0.8314	**0.8145**	0.8227	0.1986	0.5794	0.5676	0.5733	**0.2345**	**0.5333**	**0.5224**	**0.5277**
RGB (R50-DC5)	0.4989	0.8144	0.8086	0.8112	0.1918	0.5511	0.5471	0.5489	0.2247	0.5099	0.5063	0.5079
RGB (R101)	0.5048	0.8256	0.8012	0.8130	0.1969	0.5653	0.5485	0.5566	0.2284	0.5215	0.5060	0.5135
RGB (R101-DC5)	0.5068	0.8233	0.8118	0.8172	0.1935	0.5564	0.5485	0.5522	0.2299	0.5172	0.5099	0.5133
RGB (Learned PE)	0.0264	0.2184	0.1863	0.1904	0.0010	0.0254	0.0215	0.0220	0.0063	0.0657	0.0559	0.0571
RED	0.4589	0.8066	0.7968	0.8013	0.1973	0.5732	0.5662	0.5694	0.2223	0.5244	0.5180	0.5209
GREEN	0.5005	0.8255	0.8093	0.8169	**0.2020**	0.5831	**0.5716**	**0.5770**	0.2319	0.5326	0.5221	0.5271
BLUE	0.4660	0.8063	0.7919	0.7986	0.1575	0.5172	0.5078	0.5121	0.2011	0.4849	0.4762	0.4802
NIR	0.3484	0.7514	0.7080	0.7287	0.0563	0.3039	0.2862	0.2946	0.1272	0.3676	0.3463	0.3564
EF (NIR Last)	0.4674	0.7959	0.7765	0.7857	0.1063	0.4075	0.3974	0.4022	0.1786	0.4267	0.4162	0.4212
EF (NIR First)	0.4638	0.8023	0.7767	0.7889	0.1075	0.4131	0.3998	0.4062	0.1759	0.4300	0.4162	0.4228
EF (RRR+NIR)	0.2699	0.6533	0.5797	0.6114	0.0351	0.2203	0.1946	0.2056	0.0879	0.2889	0.2559	0.2701
EF (GGG+NIR)	0.2888	0.6386	0.5760	0.6031	0.0358	0.2153	0.1936	0.2030	0.0910	0.2810	0.2532	0.2652
EF (BBB+NIR)	0.1791	0.4977	0.4192	0.4494	0.0215	0.1522	0.1268	0.1366	0.0585	0.2112	0.1771	0.1902
EF (R+NIR)	0.3559	0.7254	0.6732	0.6971	0.0577	0.3027	0.2806	0.2907	0.1184	0.3494	0.3240	0.3356
EF (G+NIR)	0.3925	0.7535	0.7201	0.7356	0.0824	0.3565	0.3404	0.3478	0.1446	0.3861	0.3688	0.3768
EF (B+NIR)	0.3553	0.7276	0.6734	0.6982	0.0585	0.3077	0.2843	0.2950	0.1206	0.3493	0.3230	0.3350
MF (Concat)	0.5160	0.8378	0.8093	0.8231	0.1906	0.5714	0.5519	0.5613	0.2292	0.5276	0.5096	0.5183
MF (Add)	**0.5168**	0.8405	0.8087	**0.8241**	0.1913	0.5731	0.5514	0.5619	0.2280	0.5252	0.5053	0.5149
MF (Attention)	0.4872	0.8338	0.8055	0.8192	0.1908	0.5776	0.5579	0.5674	0.2237	0.5295	0.5115	0.5202
MF (RRR+NIR)	0.4617	0.8164	0.8015	0.8086	0.1623	0.5399	0.5300	0.5347	0.2055	0.5001	0.4909	0.4953
MF (GGG+NIR)	0.4903	0.8167	0.8036	0.8098	0.1666	0.5475	0.5387	0.5429	0.2150	0.5055	0.4974	0.5013
MF (BBB+NIR)	0.4535	0.8081	0.7922	0.7998	0.1540	0.5099	0.4998	0.5046	0.1983	0.4818	0.4722	0.4768
RGB (RES 90%)	0.5159	0.8562	0.7853	0.8190	0.1871	**0.5942**	0.5449	0.5683	0.2311	0.5331	0.4889	0.5098
RGB (RES 80%)	0.4102	**0.8624**	0.6751	0.7571	0.1385	0.5240	0.4101	0.4600	0.1763	0.4999	0.3913	0.4389
RGB (RES 70%)	0.2832	0.8494	0.4985	0.6281	0.0689	0.4342	0.2547	0.3210	0.1075	0.4575	0.2685	0.3383
RGB (RES 60%)	0.1490	0.7569	0.3544	0.4825	0.0247	0.2811	0.1315	0.1791	0.0552	0.3561	0.1667	0.2270
RGB (RES 50%)	0.0695	0.6419	0.2198	0.3273	0.0067	0.1558	0.0533	0.0795	0.0234	0.2542	0.0870	0.1296

Table 8: Class-wise evaluation metrics for DETR-based models at IoU = 0.50. Precision, recall, and F1 score are reported separately for each target class: Healthy Plant, Underdeveloped Plant, and Missing Plant. These metrics show how well each model detects each anomaly type and healthy plants at this commonly used IoU threshold.

Model	**Healthy Plant**			**Underdeveloped Plant**			**Missing Plant**		
	P	R	F1	P	R	F1	P	R	F1
RGB (R50)	0.9336	0.8680	**0.8996**	0.6997	0.4632	0.5574	0.6126	0.6246	0.6185
RGB (R101)	0.9317	0.8500	0.8890	0.7019	0.4117	0.5189	0.6000	0.6074	0.6037
RGB (R101-DC5)	0.9297	0.8515	0.8889	0.6948	0.4552	0.5500	0.6448	**0.6483**	**0.6465**
RGB (Learned PE)	0.2707	0.0481	0.0817	-	0.0000	0.0000	-	0.0000	0.0000
RED	0.9293	0.8629	0.8949	0.5960	0.4399	0.5062	0.5770	0.5532	0.5649
GREEN	0.9390	0.8611	0.8984	0.6465	0.4589	0.5368	0.6352	0.6149	0.6249
BLUE	0.9277	0.8436	0.8837	0.6037	0.3840	0.4694	0.6203	0.5865	0.6029
NIR	0.9008	0.7825	0.8375	0.4792	0.0779	0.1340	0.4442	0.4766	0.4598
EF (NIR Last)	0.9085	0.8194	0.8616	0.5910	0.3626	0.4494	0.6429	0.6116	0.6269
EF (NIR First)	0.9162	0.8216	0.8663	0.5726	0.3288	0.4178	0.6315	0.6051	0.6180
EF (RRR+NIR)	0.8973	0.5723	0.6988	0.3637	0.1834	0.2439	0.3544	0.2811	0.3135
EF (GGG+NIR)	0.8989	0.5328	0.6691	0.3045	0.2877	0.2959	0.3702	0.3362	0.3524
EF (BBB+NIR)	0.8638	0.3658	0.5140	0.2413	0.1706	0.1999	0.2696	0.2261	0.2459
MF (Concat)	0.9207	0.8666	0.8928	0.7281	**0.4896**	**0.5855**	**0.6710**	0.6081	0.6380
MF (Add)	0.9246	0.8722	0.8976	0.7458	0.4589	0.5682	0.6451	0.5861	0.6142
MF (Attention)	0.9204	**0.8745**	0.8968	0.7256	0.4736	0.5731	0.6320	0.5614	0.5946
MF (RRR+NIR)	0.9305	0.8666	0.8974	0.5597	0.4055	0.4703	0.5998	0.5722	0.5857
MF (GGG+NIR)	0.9264	0.8585	0.8912	0.6248	0.4270	0.5073	0.6116	0.6004	0.6059
MF (BBB+NIR)	0.9237	0.8559	0.8885	0.5291	0.3957	0.4528	0.5971	0.5668	0.5815
RGB (R50-DC5)	0.9284	0.8552	0.8903	0.6993	0.4423	0.5419	0.5926	0.6159	0.6040
EF (R+NIR)	0.9009	0.6891	0.7809	0.3903	0.1920	0.2574	0.5401	0.4992	0.5189
EF (G+NIR)	0.9078	0.7385	0.8144	0.4236	0.2773	0.3352	0.5917	0.5614	0.5761
EF (B+NIR)	0.9037	0.6802	0.7762	0.3365	0.2147	0.2622	0.5550	0.4896	0.5202
RGB (RES 90%)	**0.9605**	0.8265	0.8885	**0.7854**	0.3797	0.5119	0.6043	0.6011	0.6027
RGB (RES 80%)	0.9537	0.7429	0.8352	0.7254	0.2929	0.4173	0.5540	0.3205	0.4061
RGB (RES 70%)	0.9010	0.5864	0.7104	0.6623	0.1192	0.2020	0.5223	0.1037	0.1730
RGB (RES 60%)	0.8074	0.4255	0.5573	0.5000	0.0022	0.0044	0.3023	0.0145	0.0277
RGB (RES 50%)	0.6754	0.2732	0.3890	-	0.0000	0.0000	0.0000	0.0000	0.0000

Table 9: Class-wise evaluation metrics for DETR-based models at IoU = 0.75. Higher IoU thresholds emphasize precise localization. Precision, recall, and F1 score are reported for each target class to evaluate model performance when stricter spatial accuracy is required.

Model	**Healthy Plant**			**Underdeveloped Plant**			**Missing Plant**		
	P	R	F1	P	R	F1	P	R	F1
RGB (R50)	0.7692	0.7152	0.7412	0.1650	0.1092	0.1314	0.0196	0.0200	0.0198
RGB (R101)	0.7548	0.6886	0.7202	0.1747	0.1025	0.1292	0.0141	0.0143	0.0142
RGB (R101-DC5)	0.7498	0.6867	0.7169	0.1657	0.1086	0.1312	0.0187	0.0188	0.0187
RGB (Learned PE)	0.0318	0.0056	0.0096	-	0.0000	0.0000	-	0.0000	0.0000
RED	0.7667	0.7119	0.7383	0.1704	0.1258	0.1447	0.0260	0.0249	0.0254
GREEN	**0.7840**	**0.7188**	**0.7500**	0.1798	**0.1276**	**0.1493**	0.0179	0.0174	0.0177
BLUE	0.7032	0.6395	0.6698	0.1196	0.0761	0.0930	0.0238	0.0225	0.0232
NIR	0.4148	0.3603	0.3857	0.1434	0.0233	0.0401	0.0144	0.0155	0.0150
EF (NIR Last)	0.5499	0.4960	0.5216	0.1080	0.0663	0.0821	0.0306	0.0291	0.0298
EF (NIR First)	0.5581	0.5005	0.5277	0.1058	0.0607	0.0772	0.0247	0.0237	0.0242
EF (RRR+NIR)	0.3619	0.2308	0.2819	0.0681	0.0344	0.0457	0.0009	0.0007	0.0008
EF (GGG+NIR)	0.3698	0.2192	0.2753	0.0578	0.0546	0.0562	0.0008	0.0007	0.0007
EF (BBB+NIR)	0.3346	0.1417	0.1991	0.0477	0.0337	0.0395	0.0011	0.0009	0.0010
MF (Concat)	0.7385	0.6951	0.7162	0.1496	0.1006	0.1203	0.0360	0.0326	0.0342
MF (Add)	0.7365	0.6948	0.7150	0.1904	0.1172	0.1451	0.0246	0.0223	0.0234
MF (Attention)	0.7377	0.7009	0.7188	0.1673	0.1092	0.1321	0.0412	0.0366	0.0388
MF (RRR+NIR)	0.7184	0.6690	0.6928	0.1262	0.0914	0.1060	0.0229	0.0218	0.0224
MF (GGG+NIR)	0.7307	0.6772	0.7029	0.1562	0.1067	0.1268	0.0256	0.0251	0.0254
MF (BBB+NIR)	0.6787	0.6289	0.6529	0.1247	0.0933	0.1067	0.0260	0.0247	0.0253
RGB (R50-DC5)	0.7473	0.6884	0.7167	0.1465	0.0926	0.1135	0.0174	0.0181	0.0177
EF (R+NIR)	0.4566	0.3492	0.3958	0.0524	0.0258	0.0345	0.0005	0.0005	0.0005
EF (G+NIR)	0.5219	0.4246	0.4683	0.0590	0.0387	0.0467	0.0101	0.0096	0.0099
EF (B+NIR)	0.4669	0.3514	0.4010	0.0365	0.0233	0.0285	0.0016	0.0014	0.0015
RGB (RES 90%)	0.7376	0.6347	0.6823	**0.1803**	0.0871	0.1175	**0.2409**	**0.2396**	**0.2402**
RGB (RES 80%)	0.6318	0.4922	0.5533	0.1503	0.0607	0.0864	0.1439	0.0832	0.1055
RGB (RES 70%)	0.4862	0.3165	0.3834	0.1039	0.0187	0.0317	0.1146	0.0228	0.0380
RGB (RES 60%)	0.3123	0.1646	0.2156	0.0000	0.0000	0.0000	0.0698	0.0033	0.0064
RGB (RES 50%)	0.1680	0.0680	0.0968	-	0.0000	0.0000	0.0000	0.0000	0.0000

Table 10: Class-wise evaluation metrics for DETR-based models at averaged IoU = 0.50:0.95. This range-based IoU evaluation provides a more comprehensive assessment of detection robustness across varying localization requirements. Precision, recall, and F1 score are reported per class.

Model	**Healthy Plant**			**Underdeveloped Plant**			**Missing Plant**		
	P	R	F1	P	R	F1	P	R	F1
RGB (R50)	**0.6579**	0.6117	0.6340	0.2955	**0.1956**	0.2354	0.1890	0.1927	0.1908
RGB (R101)	0.6484	0.5915	0.6186	0.2868	0.1682	0.2121	0.1765	0.1787	0.1776
RGB (R101-DC5)	0.6459	0.5916	0.6176	0.2824	0.1850	0.2236	0.1938	0.1949	0.1944
RGB (Learned PE)	0.0825	0.0147	0.0249	-	0.0000	0.0000	-	0.0000	0.0000
RED	0.6571	0.6101	0.6328	0.2524	0.1863	0.2143	0.1870	0.1792	0.1830
GREEN	0.6681	**0.6126**	**0.6392**	0.2694	0.1912	0.2237	0.1890	0.1830	0.1860
BLUE	0.6110	0.5556	0.5820	0.2272	0.1445	0.1767	0.1919	0.1814	0.1865
NIR	0.4671	0.4057	0.4343	0.2125	0.0345	0.0594	0.1333	0.1430	0.1380
EF (NIR Last)	0.5270	0.4753	0.4998	0.2117	0.1299	0.1610	0.2085	0.1984	0.2033
EF (NIR First)	0.5318	0.4768	0.5028	0.2078	0.1193	0.1516	0.1987	0.1904	0.1945
EF (RRR+NIR)	0.4369	0.2787	0.3403	0.1302	0.0656	0.0873	0.0646	0.0512	0.0571
EF (GGG+NIR)	0.4428	0.2625	0.3296	0.1119	0.1058	0.1088	0.0580	0.0526	0.0552
EF (BBB+NIR)	0.4130	0.1749	0.2458	0.0898	0.0634	0.0743	0.0524	0.0440	0.0478
MF (Concat)	0.6340	0.5968	0.6148	0.2992	0.2012	0.2406	0.2139	0.1938	0.2034
MF (Add)	0.6282	0.5926	0.6099	0.3167	0.1949	**0.2413**	0.2065	0.1876	0.1966
MF (Attention)	0.6333	0.6018	0.6172	0.2903	0.1895	0.2293	0.2107	0.1872	0.1982
MF (RRR+NIR)	0.6203	0.5777	0.5982	0.2192	0.1588	0.1842	0.1883	0.1796	0.1838
MF (GGG+NIR)	0.6266	0.5807	0.6028	0.2606	0.1781	0.2116	0.1937	0.1902	0.1919
MF (BBB+NIR)	0.5967	0.5529	0.5740	0.2118	0.1584	0.1813	0.1887	0.1791	0.1838
RGB (R50-DC5)	0.6416	0.5911	0.6153	0.2773	0.1754	0.2149	0.1764	0.1833	0.1798
EF (R+NIR)	0.4845	0.3706	0.4199	0.1261	0.0620	0.0831	0.1113	0.1029	0.1069
EF (G+NIR)	0.5124	0.4169	0.4598	0.1464	0.0958	0.1158	0.1648	0.1564	0.1605
EF (B+NIR)	0.4858	0.3657	0.4173	0.1068	0.0682	0.0832	0.1141	0.1006	0.1069
RGB (RES 90%)	0.6314	0.5433	0.5840	**0.3206**	0.1550	0.2090	**0.2918**	**0.2903**	**0.2910**
RGB (RES 80%)	0.5799	0.4517	0.5079	0.2663	0.1075	0.1532	0.2205	0.1276	0.1616
RGB (RES 70%)	0.4986	0.3245	0.3931	0.2377	0.0428	0.0725	0.1975	0.0392	0.0654
RGB (RES 60%)	0.3882	0.2046	0.2679	0.1000	0.0004	0.0009	0.1023	0.0049	0.0094
RGB (RES 50%)	0.2699	0.1092	0.1555	-	0.0000	0.0000	0.0000	0.0000	0.0000