American Society of Agricultural and Biological Engineers
2950 Niles Road | St. Joseph MI 49085-9659 | USA
269.429.0300 | fax 269.429.3852 | hq@asabe.org | www.asabe.org

An ASABE Meeting Presentation
DOI: https://doi.org/10.13031/aim.202400615
Paper Number: 2400615

Robotic Plot-scale Peanut Counting and Yield Estimation using LoFTR-based Image Stitching and Improved RT-DETR

Zhengkun Li[1], Rui Xu[1], Changying Li[1], Barry Tillman[2,3], Nino Brown[4]

[1] Bio-Sensing, Automation, and Intelligence Lab, Department of Agricultural and Biological Engineering, University of Florida, Gainesville, Florida, USA
[2] North Florida Research and Education Center, University of Florida, Marianna, Florida, USA
[3] Agronomy Department, University of Florida, Gainesville, FL, USA
[4] Department of Crop and Soil Science, University of Georgia, Tifton, GA, USA

Written for presentation at the
2024 ASABE Annual International Meeting
Sponsored by ASABE
Anaheim, CA
July 28-31, 2024

ABSTRACT. *Peanuts, ranking as the seventh-largest crop in the United States with a farm value exceeding $1 billion, are pivotal to global food security. Conventional peanut yield estimation methods involve digging, harvesting, transporting, and weighing, which are labor-intensive and inefficient for large-scale operations. This inefficiency is particularly pronounced in peanut breeding, where requires precise yield estimations of each plot-scale pods for genotypes comparison and selection. We proposed an automated approach utilizing a robotic system equipped with machine vision to predict peanut yields post-digging and inverting. This system leverages a mobile robot with an imaging system that captures sequential images of peanut plots, each representing a different genotype, utilizing spatial geographic information. A robust hierarchical strategy was introduced for plot-scale image stitching, employing a Local Feature Transformer (LoFTR)-based feature matching algorithm. Additionally, the Real-Time Detection Transformer (RT-DETR) was customized for pod detection by integrating partial convolution into a lightweight ResNet-18 backbone and refining the upsampling and downsampling modules in Cross-scale Feature Fusion. Our methods were validated in two breeding fields, where the LoFTR-based stitching achieved approximately three times denser and more uniform feature matching than the conventional Scale-Invariant Feature Transform (SIFT) approach. The customized peanut pod detector demonstrated a mean Average Precision (mAP50) of 89.3% and an mAP95 of 55.0% with lighter weights and less computation, improving by 3.3% and 5.9%, respectively, over the original RT-DETR model. Finally, we deployed the detector on the stitched plot-scale images and calculated the pods number for predicting the yield. Achieving a Mean Absolute Percentage Error (MAPE) of 9% and an R-square of 0.47, our approach outperforms the mainstream Structure from Motion (SfM) based methods. This innovative approach significantly reduces the time and labor required for yield determination, thereby advancing the efficiency of peanut breeding operations in complex, dynamic outdoor environments.*

Keywords. *Peanut, High-Throughput Phenotyping, Yield estimation, Image stitching, Pod detection.*

1. Introduction

Peanuts, recognized as the seventh-largest crop in the United States with a valuation exceeding $1 billion, play a crucial role in the national economy and global food security (Tellus, 2021). As demand for peanuts grows, the field of peanut agriculture increasingly incorporates technological innovations to enhance the efficiency and precision of yield prediction and breeding processes.

Historically, peanut breeding has focused on improving yield, disease resistance, and environmental adaptability. The adoption of modern genetic and agronomic techniques has transformed breeding programs, allowing for the rapid selection of high-performing genotypes. Recent advancements have prominently featured the use of unmanned aerial vehicles (UAVs) for high-throughput phenotyping, which is pivotal in accelerating breeding cycles and enhancing crop traits. This approach has diverse applications in peanut cultivation, ranging from variety selection (Balota & Oakes, 2016) to disease resistance phenotyping (Patrick et al., 2017), as well as advanced assessments of canopy health, including leaf disease detection (Larsen et al., 2022), canopy height measurement (Sarkar et al., 2020), leaf area index evaluation (Sarkar, Cazenave, et al., 2021) and leaf wilting analysis (Sarkar, Ramsey, et al., 2021). The incorporation of multispectral and RGB imaging in UAVs, as demonstrated by several research groups (Bagherian et al., 2023; Manley et al., 2023) has further expanded the capabilities of UAV-based phenotyping. These advancements allow for detailed evaluation of physiological traits under stress conditions and provide avenues for predicting yields in peanut farming. These studies highlight the efficacy of UAVs in capturing a broad spectrum of phenotypic data, which is essential for informed breeding and management decisions.

Despite the broad applications of UAVs, ground-based systems offer distinct advantages, particularly in terms of data accuracy and resolution. Ground-based systems can provide detailed, close-range data capture that allows for finer resolution images and measurements of the peanut canopy architecture (Yuan et al., 2019; Yuan et al., 2018). This level of detail is critical for assessing micro-variations in crop growth and health, which can be obscured by the higher-altitude perspectives of UAVs. The use of sophisticated sensors and non-invasive technologies like ground-penetrating radar further enhances the ability of ground-based systems to assess yields and subsurface characteristics without disturbing the crop architecture (Dobreva et al., 2021).

However, since peanut pods develop underground, existing methods struggle to directly detect pods and accurately assess their growth status. These methods generally resort to rough regression and estimation for the yield evaluation, which is compromised by outdoor environmental factors such as weather, varying light conditions, genotype differences, and sensor precision. Consequently, these approaches lack the reliability and confidence required to replace traditional post-harvest field operations such as digging, inverting, harvesting, weighing, and yield assessment. One potential solution is using a robotic imaging system to directly detect peanut pods to acquire real growing situation of peanut pods after the digging and inverting processes in the field. Puhl's group developed a deep neural network approach for infield peanut pod counting, demonstrating the potential synergy between advanced imaging techniques and machine learning to enhance the accuracy of yield estimation (Puhl et al., 2021). But their system estimates the entire plot's condition by collecting videos and sampling each plot (genotype). Given the considerable spatial variability in peanut pod distribution within a plot, influenced by the quality of sampling, this method can lead to significant discrepancies in the final plot-scale pod count and yield estimation.

Inspired by existing research, we proposed to develop an algorithm capable of combining a sequence of images collected from a single peanut plot in the field into a panoramic-like long image using image stitching techniques. A deep learning model was employed to detect pods within this stitched image, enabling accurate estimation of the entire plot-scale pod level. This approach was designed to mitigate issues related to the spatial variability in pod distribution that typically affects sampling quality and yield estimations. By creating a comprehensive view of the entire plot, our method aimed to provide a more accurate and consistent assessment of pod counts and potential yield, reducing the risk of discrepancies that arise from partial or uneven sampling.

Specifically, our research employed a mobile robotic system equipped with an advanced imaging system, designed to autonomously navigate through peanut fields and capture high-resolution image sequences. Utilizing GPS and sophisticated sensors for precise positioning, the robot systematically recorded data that are processed using a robust hierarchical image stitching algorithm based on the Local Feature Transformer (LoFTR) (Sun et al., 2021). This method ensures precise alignment and synthesis of extensive image data, essential for comprehensive plot analysis. Further, we refined the Real-Time Detection Transformer (RT-DETR) (Lv et al., 2023) equipped with a lightweight ResNet18 backbone, enabling efficient peanut pod detection from the stitched images. This setup was particularly designed to handle variations in pod size, shape, and occlusion, utilizing advanced machine learning techniques to enhance detection accuracy.

The primary objectives of our research are 1) developing a robust image stitching algorithm utilizing a LoFTR-based matching algorithm for generating plot-scale images, 2) customizing the RT-DETR detector specifically for the detection of peanut pods, and 3) evaluating the accuracy of plot-scale pods counts and yield with our proposed analytical approach.

2. Materials and Methods

This study focused on developing a robotic vision system that is capable of realizing plot-scale peanut pod counting for the yield estimation. The overall procedure is presented in the flowchart shown in Figure 1. In the data collection stage, we modified the modular agricultural robot, MARS-X (Xu & Li, 2022), to adapt navigating in the row-type peanut breeding field. The image sequences were recorded during the robot scanning the peanut field and the images of individual peanut plot were extracted based on the geographic information. The collected images were then processed through a robust stitching algorithm to collectively generate a panoramic-like, plot-scale image, including the steps of LoFTR Feature Matching, RANSAC Homography Estimation, and a Hierarchical Stitching Strategy. Meanwhile, a pods detection dataset was used to train a neural network-based pods detector, specifically, we customized the state-of-the-art RT-DETR detector to enhance the pods detection. Finally, a sliced inference strategy, Slicing Aided Hyper Inference (SAHI) (Akyon et al., 2022), was applied to the high-resolution stitched images that applied the detector on subdividing slices to enhance the small pods detection. The pipeline enables to output pod number of individual peanut plot as results to compare with the real peanut pod's number and yield for pipeline evaluation.

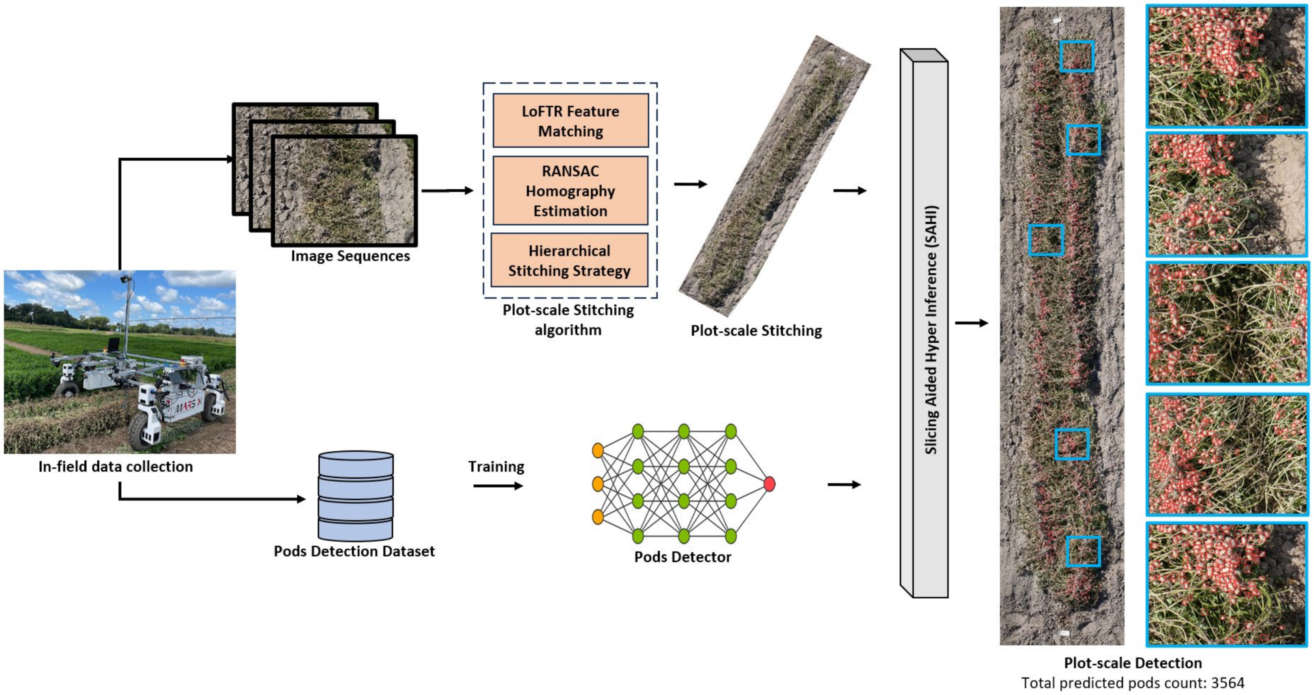


Image Sequences
LoFTR Feature Matching
RANSAC Homography Estimation
Hierarchical Stitching Strategy
Plot-scale Stitching algorithm
Plot-scale Stitching
In-field data collection
Training
Pods Detection Dataset
Pods Detector
Slicing Aided Hyper Inference (SAHI)
Plot-scale Detection
Total predicted pods count: 3564

Figure 1. Workflow of automated plot-scale peanut pod counting using image stitching and deep learning detection. A Transformer-based feature matching algorithm was used for robust plot-scale stitching with image sequences. Then a pod detector was applied to detect pods on stitched image with sliced inference.

2.1 Data acquisition and preparation

Two peanut breeding fields were involved in the project: IFAS Plant Science Research and Education Unit (PSREU, Citra, FL, USA) and Georgia Tifton peanut breeding field (Tifton, GA, USA). We used different robotic platforms and different camera configuration to capture the images sequences or videos in the two field to enhance the diversity of data source. In the case of Citra field (Figure 2a), the robotic platform MARS-X (Xu & Li, 2022) equipped with three types of cameras to acquire the in-field imaging data: Panasonic Lumix G7 mirrorless cameras (Panasonic, Osaka, Japanese), Raspberry Pi high quality 3MP cameras (Raspberry Pi Foundation, Cambridge, United Kingdom), and RealSense D435i camera (Intel RealSense, California, United States). In Tifton's field, another mobile platform "Watson", used the FLIR Blackfly cameras (Teledyne FLIR, Oregon, United States) to capture the imaging data with a larger field of view (Figure 2b). The two robots navigated in the field with Dual-GNSS system over the digging peanut plots. They moved with a constant linear velocity of 0.3m/s and recorded the image or videos at the same time.

The practical challenges in field data collection are manifold, often influenced by variable field conditions. These include fluctuations in natural light due to weather changes or time of day, occlusions caused by pod clustering, and shadows from surrounding plants and objects (Figure 2c and 2d). Additionally, the narrow window for data collection that is only available after digging and before harvesting, increasing the complexity of obtaining reliable image data from peanut pods during the growing season.

Two types of datasets including the plot-scale dataset and peanut pod detection dataset were built for the plot-scale stitching and plot-scale pod counting. Specifically, plot-scale dataset contains the images sequences of single plots according to the geographic information. The dataset only contains the high-resolution images from Lumix G7 and FLIR Blackfly cameras, aiming to generate the high-quality panorama-like plot-scale images. While the peanut pod detection dataset contains the images and their manual annotations from all the camaras for training a more robust detector in different configurations (Table 1). The detection dataset mixed the annotated data from the two fields and then was divided into training and valid dataset with the ratio of 7:3. Before model training, the training data was augmented into three times with random brightness, blur, cropping and mosaic operations.

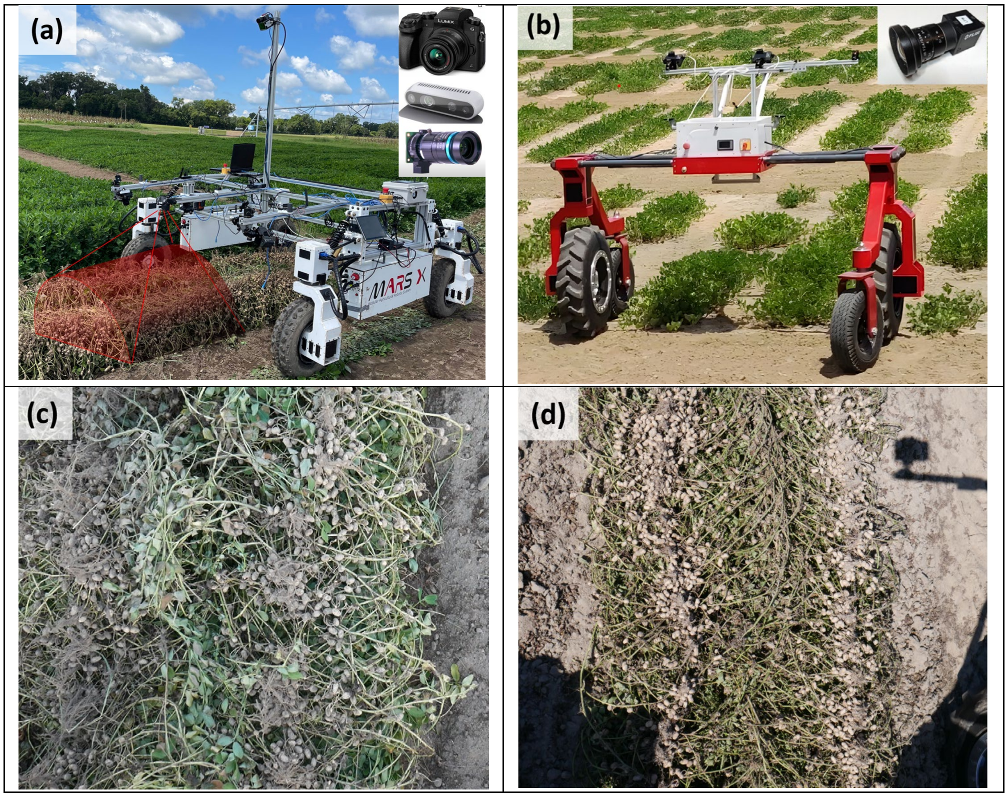


Figure 2. In-field Robotic platform and data acquisition. (a) Robotic imaging system in the peanut field; (b) The scanning field in Citra; (c) and (d) are the two examples of typical images with/without shadows.

Table.1 Dataset preparation and preparation

Field	Time	Data type	Device	# Plot sequences	Pod Detection Dataset	
					# image	# instance
Citra	10/23/2023	Video (1080P, 30fps) Image (4k, 1fps)	Lumix G7 Pi cam, RS D435i	150	188	32,618
Tifton	10/18/2023	Image (4k, 1fps)	Point Gray	60	245	29,627

2.2 Hierarchical stitching strategies based on LoFTR

In the pursuit of generating comprehensive plot-scale images for peanut pod detection, we employed a hierarchical stitching strategy based on the Local Feature Transformer (LoFTR). In contrast to the traditional sequential stitching process, where each image is added one after the other, potentially amplifying errors with each step, the hierarchical method treats image stitching as a structured, tiered process. This allows for correction of errors at lower levels of the hierarchy, preventing them from propagating through the entire dataset.

The flowchart of hierarchical stitching approach is shown in the Figure 3, which leverages the strengths of LoFTR to achieve high-quality panoramic-like image synthesis. For each plot $\boldsymbol{i}$, this strategy begins with the individual images $\boldsymbol{I_1^i}$ of the plot, then pairs of images $\boldsymbol{I_{j:j+1}^i}$ are processed through feature matching based on a LoFTR-based description, resulting in a series of stitched images $\boldsymbol{I_{j:j+2}^i}$ and $\boldsymbol{I_{j:j+4}^i}$ and so on, until the entire sequence culminates in a comprehensive plot-scale image $\boldsymbol{I_{1:m}^i}$. This hierarchical method markedly reduces the compound errors that typically accumulate in one-by-one stitching processes. By matching features in overlapping image pairs and progressively merging them into larger composites, the approach minimizes discrepancies at each level, ensuring that the final panoramic image retains geometric and photometric consistency.

Traditional feature-based methods like SIFT (Scale-Invariant Feature Transform) and ORB (Oriented FAST and Rotated BRIEF) often encounter challenges in agricultural fields, where the scenery may exhibit repetitive patterns, similar-looking vegetation, and varying lighting conditions, leading to insufficient or incorrect feature correspondences. The adoption of the Local Feature Transformer (LoFTR) algorithm (Sun et al., 2021) offers a significant advancement in addressing these challenges. Unlike traditional feature-based methods, LoFTR operates on a detector-free mechanism, leveraging the power of Transformer architecture to establish dense and context-aware feature correspondences. This approach enables LoFTR to effectively handle the repetitive and low-texture regions commonly found in peanut fields, ensuring precise image alignment. Random sample consensus (RANSAC) was then utilized for homography estimation to filter out outlier correspondences and ensure accurate alignment.

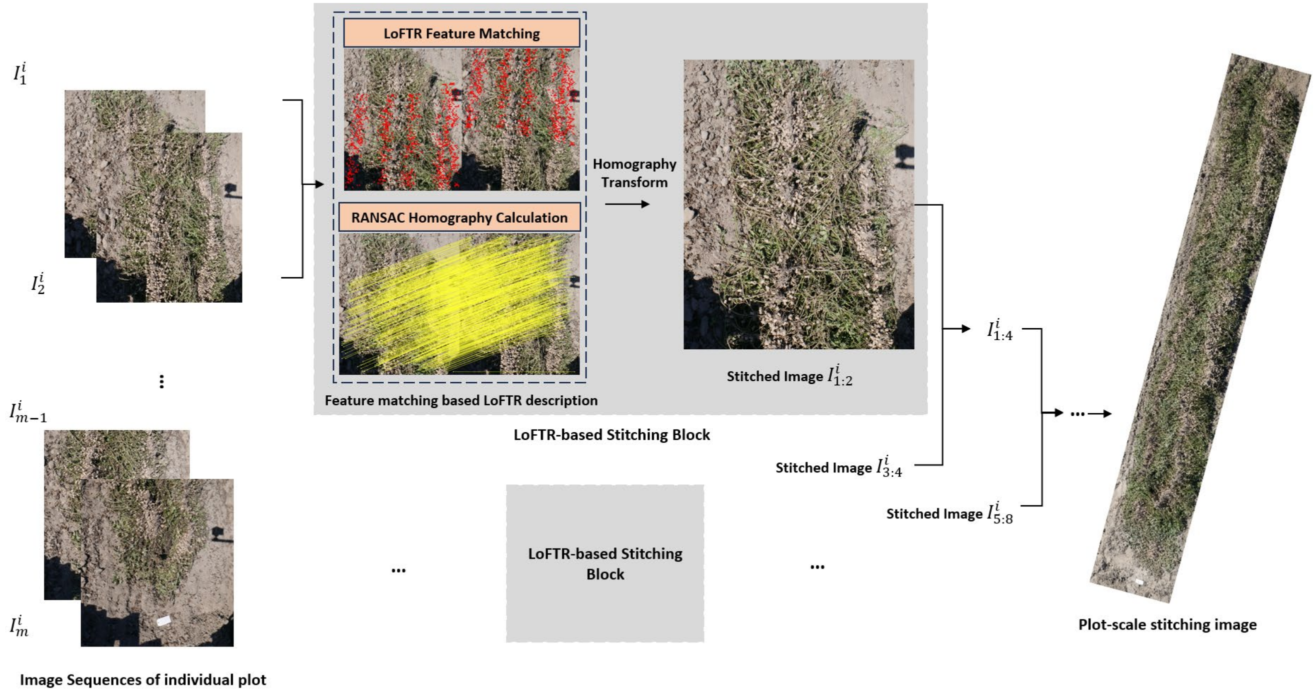

I_1^i
I_2^i
⋮
I_{m-1}^i
I_m^i
Image Sequences of individual plot
LoFTR Feature Matching
RANSAC Homography Calculation
Feature matching based LoFTR description
Homography Transform
Stitched Image $I_{1:2}^i$
LoFTR-based Stitching Block
$I_{1:4}^i$
Stitched Image $I_{3:4}^i$
Stitched Image $I_{5:8}^i$
···
···
LoFTR-based Stitching Block
···
···
Plot-scale stitching image

Figure 3: Hierarchical strategy for plot-scale image stitching that divide images into stitching groups rather than stitching by sequence. In the LoFTR-based stitching block, the red dots present the matching correspondences between two images and the yellow line is linked the inlier matches after RANSAC filtering.

2.3 Improved RT-DETR for peanut pod detection

Real-Time Detection Transformer (RT-DETR) introduced by Lv et al. (2023) advances the domain of object detection by integrating Vision Transformers (ViT) for real-time, multiscale feature processing, outperforming many traditional detectors in speed and accuracy. Unlike other real-time detectors such as YOLO, RT-DETR circumvents the need for post-processing like NMS, streamlining inference. Its architecture includes a versatile CNN backbone and an innovative hybrid encoder that decouples interactions within and across scales, enhancing feature utilization. A novel IoU-aware query selection mechanism improves the initialization of object queries, facilitating precise detection.

To suit the computational constraints of robotic systems in the field, we have tailored the RT-DETR model with a ResNet18 backbone, optimizing it for lightweight, on-board computation without compromising real-time performance (Figure 4a). This adaptation allows for agile adjustment of the decoding layers to balance detection accuracy with processing speed, negating the need for retraining. Our custom model capably meets the demands of in-situ peanut pod detection, combining efficiency with the robust capabilities of RT-DETR.

Specifically, inspiring by FasterNet (Chen et al., 2023), the partial convolutional technology was applied to build the partial convolutional layer (Pconv) in the basic block of ResNet to reduce the number of floating-point operations and parameters, thereby streamlining the model's complexity. A customized FasterBlock was designed to replace the original resent block. After the feature extraction with ResNet18-FasterBlock backbone, the features from the last three stages of the backbone were fed into the encoder. The efficient hybrid encoder transforms multi-scale features into a sequence of image features through the Attention-based Intra-scale Feature Interaction (AIFI) and the CNN-based Cross-scale Feature Fusion (CCFF). Two state-of-the-art modules of up sampling (DySample) (Liu et al., 2023) and down sampling (ADown) (Wang et al., 2024) were applied to enhance the feature fusion process in CCFF. Then, the uncertainty-minimal query selection selects a fixed number of encoders features to serve as initial object queries for the decoder. Finally, the decoder with auxiliary prediction heads iteratively optimizes object queries to generate categories and boxes.

ResNet18-FasterBlock (Backbone): As illustrated in Figure 4b, the FasterBlock module evolves from the original ResNet block by substituting its traditional convolutional layer with the partial convolution layer (Chen et al., 2023), which has been proved markedly enhances the efficiency of feature extraction. The block is particularly advantageous for smaller network architectures, as it contributes to faster execution speeds. This is especially beneficial in resource-constrained environments such as embedded systems and mobile devices.

A FasterBlock is comprising a PConv followed by two 1 × 1 Conv layers as a reverse residual block, the FasterNet Block also incorporates shortcut connections to reuse important features. The normalization layer and activation layer are only used after the middle 1 × 1 Conv layer, which plays the role of retaining feature diversity and realizing lower delay. In this process, PConv selectively utilizes a subset of the input channels for spatial feature extraction through regular convolutions, while the remaining channels are left unaltered. This approach not only streamlines the feature extraction process but also contributes to the overall efficiency of the neural network. Assuming the input and output feature maps have an equal number of channels, denoted as c, the FLOPs of PConv are $h \times \mathrm{w} \times k^2 \times c_p^2$. When the partial ratio $r = c_p/c = 1/4$, the FLOPs of PConv are only $1/16$ of ordinary convolution. Meanwhile, the memory access amount of PConv is calculated as $h \times \mathrm{w} \times k^2 \times c_p^2 \approx h \times \mathrm{w} \times 2c_p^2$, which is about $1/4$ of the original.

DySample (Upsampling in CCFF): DySample is a high-efficient dynamic feature sampling module (Liu et al., 2023) in dense prediction models for gradually recovering the feature resolution. It redefines the upsampling task as dynamic point re-sampling rather than relying on fixed kernels or convolution operations. Traditional methods like nearest-neighbor or bilinear interpolation, apply predefined rules that don't consider the semantic content of the feature map, leading to less than optimal upsampling results. DySample utilizes a feature map's spatial information to adaptively generate sampling points, thus allowing for a more content-aware upsampling. The system dynamically computes offsets based on the low-resolution input and applies these to a higher-resolution grid, effectively resampling points according to the underlying semantic structure of the feature map. This strategy allows the upsampler to adapt to different areas of the image, preserving detail and reducing artifacts.

The sampling process based dynamic upsampling is shown in Figure 4c. Given a feature map χ of size $C \times H \times W$ and a sampling set S of size $2g \times H \times W$, where 2 of the first dimension denotes the x and y coordinates and g is a hiperparapmters of sampling group number. the grid sample function uses the positions in S to re-sample the hypothetical bilinear-interpolated χ into χ' of size $C \times sH \times sW$ with the upsampling scale factor s. This process is defined by:

$$\chi' = grid\_sample(\chi, S) \tag{1}$$

A grouped 2D convolutional layer with the group number of g was used for updating the dynamic offset $\mathcal{O}$ of size $2s^2 \times H \times W$, whose input and output channel numbers are C and $2gs^2$. It is then reshaped to $2 \times sH \times sW$ by Pixel Shuffling (Shi et al., 2016). Then the sampling set S is the sum of the offset $\mathcal{O}$ and the original sampling grid $\mathcal{G}$, i.e.,

$$\mathcal{O} = Pixel\_shuffle\ (conv2d(\chi)) \tag{2}$$

$$S = \mathcal{G} + \mathcal{O} \tag{3}$$

where the reshaping operation is omitted. Finally, the upsampled feature map χ' of size $C \times sH \times sW$ can be generated with the sampling set by grid sample as Eq. (1). In the Implementation, we choose the scale factor s with 2 and group number with 4.

ADown (Downsampling in CCFF): ADown module was first proposed in YOLOv9 (Wang et al., 2024), specifically designed to enhance the downsampling process in convolutional neural networks (CNNs). By integrating both average and max pooling layers prior to convolutional operations, this module can capture an array of features that encapsulate both the average texture of the input data and its most prominent details (Figure 4d). This dual-pathway approach ensures that the resulting feature maps are rich and robust, containing a mix of general and specific details that may be critical for the task at hand. Moreover, the average pooling component contributes to noise reduction, potentially increasing the network's robustness to variations in input data. The concatenation of features from diverse pooling methods allows the network to maintain a comprehensive understanding of the input, which could be crucial for complex tasks like image segmentation or object recognition.

In contrast to traditional convolutional layers that employ striding for downsampling, the ADown module may present an enhanced efficiency in parameter usage and computational cost. While stride convolutions reduce spatial dimensions, they can inadvertently discard valuable information. The "ADown" module could sidestep the pitfall by initially compressing inputs in a manner that conserves different types of information. Additionally, pooling operations require no learnable parameters, which can lead to a more parameter-efficient design. The module's configuration could also offer computational advantages; by reducing dimensions early on via pooling, it sets the stage for faster subsequent convolutions. This strategic reduction and expansion of feature sets not only enables a deeper network to learn a hierarchy of features but also provide flexibility to adapt to different data types and tasks, potentially resulting in improved learning outcomes.

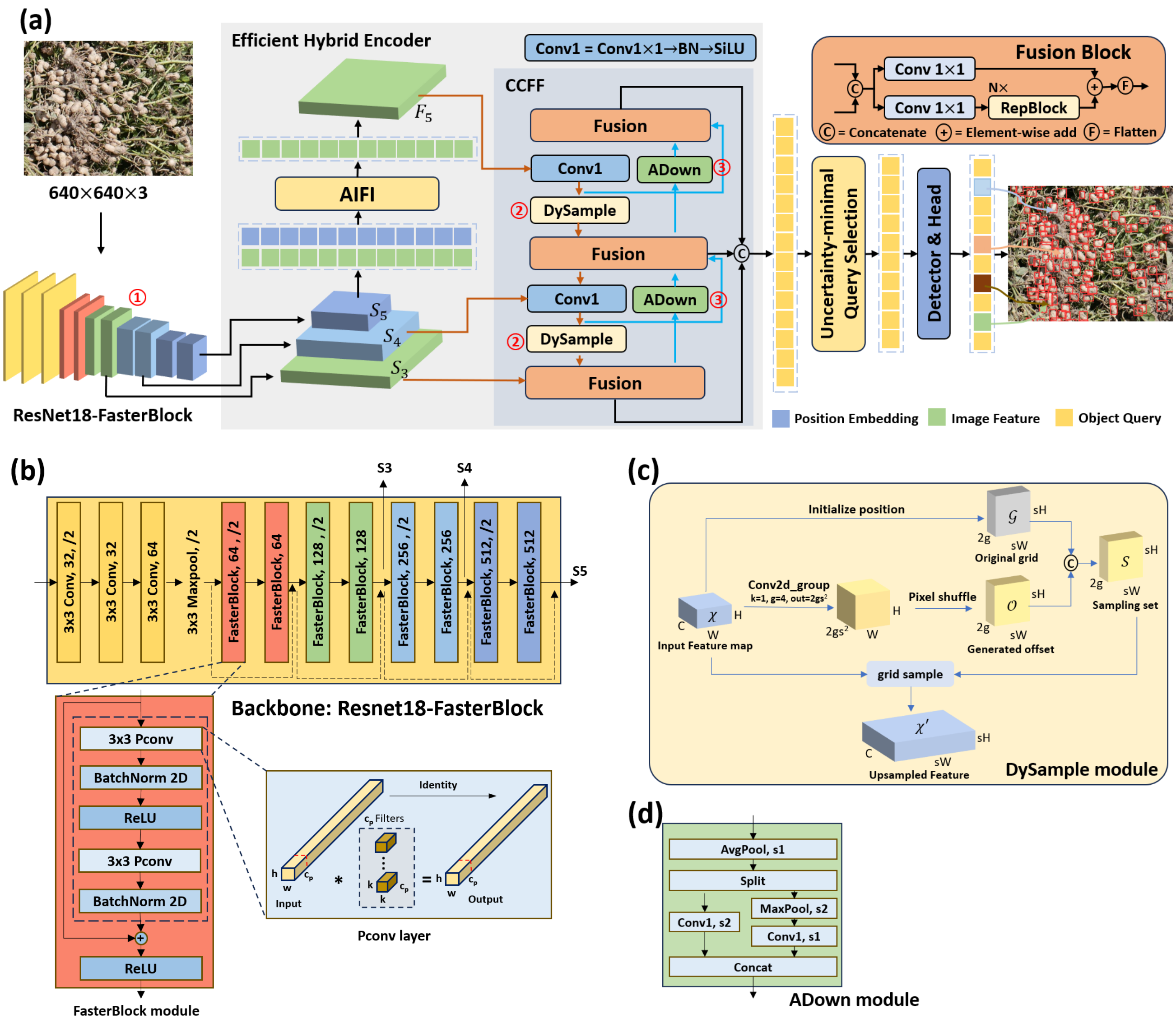


Figure 4. Illustration of improved RT-DETR detector. (a) overview of customized RT-DETR detector (Lv et al., 2023): the block of ①,②,③ highlight the three improvements: faster resnet18-FasterBlock backbone, learnable dynamic sampler DySample, and a high-efficient down sapling module Adown; (b) Backbone of ResNet18-FasterBlock; (c) Up sampling based on DySample (Liu et al., 2023); (d) Adown module for down sampling (Wang et al., 2024).

3. Experiments and Results

3.1 Performance of LoFTR-based matching

In order to improve the precision of homography estimations within the domain of in-field peanut imagery, we conducted an empirical analysis comparing the efficacy of different distinct feature detection algorithms: Scale-Invariant Feature Transform (SIFT), Oriented FAST and Rotated BRIEF (ORB), and the Local Feature Transformer (LoFTR). As shown in

Figure 5, the experimental workflow is aimed at evaluating the accuracy of different feature matching algorithms in our in-field scenarios without the image pairs and their corresponding homography matrix to serve as a ground truth. We simulate transformations by applying random homographies and various changes to the original image, creating pairs of images (original and transformed). Firstly, random augmentations were processed to increase the variations from the original image including the brightness and contrast adjustments, blur, and more importantly local shadow operation, which is simulate the real in-field situation. Then we generate a random homography H as the "ground truth" to transform the augmented image that contains transition, rotation, scale, shear and perspective. In this case, we can get the referenced image pairs and their homography matrix. By using different feature matching algorithms like LoFTR, SIFT, and ORB, we estimate the homography matrix H' for the constructed image pairs, and perform a warp transformation on the augmented images to obtain the wrapped images. Thus, we can assess the relative performance of different algorithms by comparing the similarity between the transformed and the wrapped images, as well as comparing the ground truth random homography with the estimated homography H'.

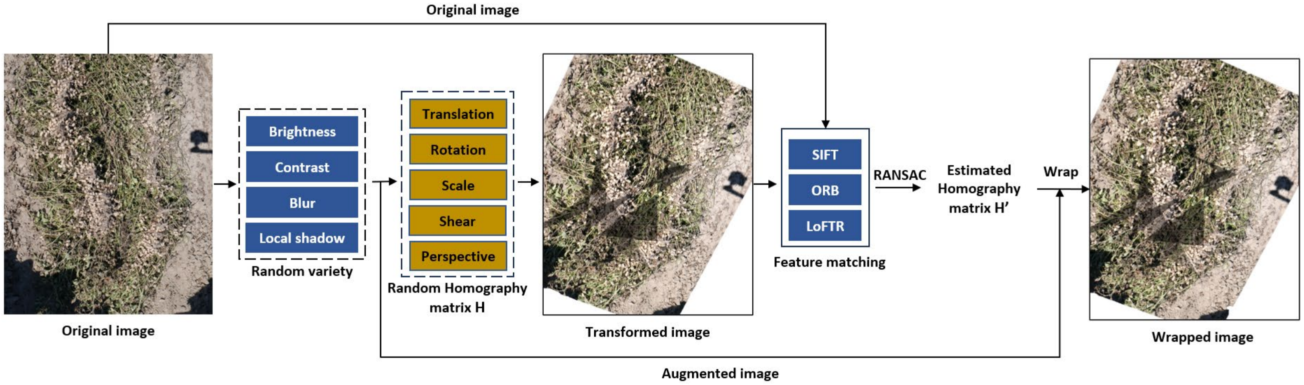


Figure 5. The comparison pipeline of different feature matching method on our in-field plot-scale dataset.

To evaluate the feature matching algorithm, we compared the image-level and matrix-level similarity metrics. Specifically, through evaluating the quality of the warped image against the original image, we aim to check how closely the warped image resembles the original, indicating the accuracy of the homography estimation.

Mean Squared Error (MSE): MSE measures the average squared difference between each pixel of the reference image and the distorted image.

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(I(i,j) - K(i,j))^2 \tag{4}$$

Where I is the reference image, K is the distorted image, and $M \times N$ is the size of the images.

Where MAX_I is the maximum possible pixel value of the image (e.g., 255 for 8-bit images).

Structural Similarity Index (SSIM): SSIM is a more sophisticated metric that considers changes in structural information, luminance, and contrast.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{5}$$

Where, μ_x, μ_y are the average intensities of images of images x and y; σ_x^2, σ_y^2 are the variances of images x and y; σ_{xy} is the covariance of images x and y. $c_1$and c_2 are constants to stabilize division with weak denominator with $c_1 = (k_1L)^2$ and $c_2 = (k_2L)^2$. L is the dynamic range of the pixel-values (usually $L = 2^{2bits\_per\_pixel} - 1$ with ($k_1 = 0.01$ and $k_2 = 0.03$ by default.

Two matrix-level similarity metrics are selected to directly compare the random homography matrix H and the estimate homography matrix H', including the metrics of Frobenius Norm Difference and Cosine similarity.

Frobenius Norm Difference: The Frobenius norm, often used to measure the size of a matrix, is analogous to the Euclidean norm for vectors. To compare two matrices of the same size, one can calculate the Frobenius norm of their difference:

$$Frobenius\_Diff(H,H') = \|H - H'\|_F = \sqrt{\sum_{i,j}(h_{ij} - h'_{ij})^2} \tag{6}$$

Cosine similarity: measures the cosine of the angle between two vectors in a multi-dimensional space. For matrices, one common approach is to flatten the matrices into vectors and then calculate the cosine similarity between these vectors:

$$Cosine\ Similarity(H,H') = \frac{vec(A) \cdot vec(B)}{\|vec(A)\|_2 \cdot \|vec(B)\|_2} \tag{7}$$

A total of 308 in-field images were randomly selected to be processed with the evaluation method. Table 2 presented the comparative performance of feature matching algorithms. Overall, LoFTR-based feature matching algorithm was better than traditional method including SIFT and ORB across most of evaluation metrics. It showed the lowest Mean Squared Error, indicating high fidelity in pixel intensity to the original image, and exceled in both Frobenius norm difference and Cosine Similarity, suggesting the highest precision in homography matrix estimation. Although LoFTR doesn't reach the peak performance of SIFT in terms of SSIM, LoFTR also maintained the structural integrity and luminance of the image most effectively. The results collectively advocated for LoFTR's robust capability, marking it as the most effective algorithm for precise feature matching in challenging in-field conditions typical of peanut pod detection tasks.

Table 2. Comparison of feature points quality among different feather matching approach

Method	MSE↓	SSIM↑	Frobenius_diff↓	Cosine Similarity↑
SIFT	460.9	**0.978**	20.63	0.952
ORB	4319.5	0.805	307.66	0.634
LoFTR	**397.2**	0.874	**12.08**	**0.992**

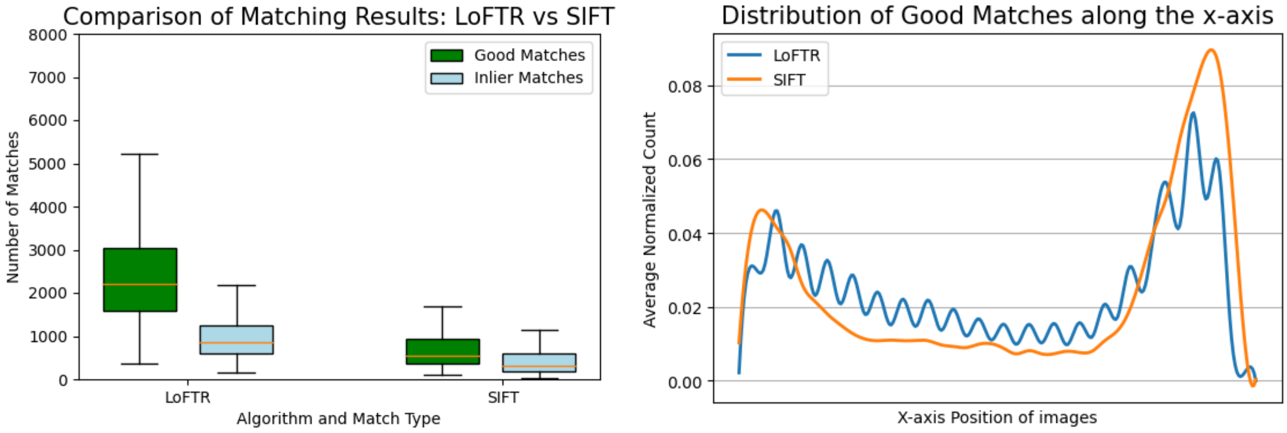
3.2 LoFTR matching for adjancent images

In assessing image stitching quality, the quantity of matching points are crucial metrics for evaluating feature matching algorithms like SIFT, ORB, and LoFTR. These metrics provide a dual assessment of both the quantity and the fidelity of feature matches. We use the following two metrics to evaluate the matching quality between two adjacent images:

# Good Matches: For SIFT and ORB, after feature matching (FLANN), a common approach is to filter out matches using the Lowe's ratio test, where only matches with a distance less than 0.75 (or some threshold) times the distance of the second-best match is kept. This can be calculated by counting the number of matches that pass this test. For LoFTR, we selected these matches with the confidence over 0.75.

# Inlier Matches: After initial matching, RANSAC (Random Sample Consensus) can be applied to filter out outliers among the matches. The matches that survive this process are usually termed 'good' or 'successful' correspondences. Count these correspondences to get the number of successful matches. Successful Correspondences Count (matches after RANSAC).

A total of 260 image pairs (adjacent two images) collected in field were randomly selected to be processed to calculate the quantity of good matches and inlier matches. LoFTR presented almost three times quantity of good matches and inlier matches than SIFT method, suggesting a more robust feature detection and matching capability (Figure 6a). Figure 6b showed the distribution of good matches along the x-axis of images, where LoFTR not only generally maintains a higher frequency of good matches across the span of the image but also exhibits a more uniform distribution. This uniformity is crucial for consistent image stitching, especially in panoramic stitching or other applications requiring alignment across wide image areas. These results collectively underscore LoFTR's superior performance in generating more reliable and evenly distributed feature matches compared to SIFT, thereby indicating its potential for improving automated image stitching tasks.



(a) (b)

Figure 6: Quantitative comparison of Image stitching between the LoFTR and SIFT. (a) quantity comparison of good matches and inlier matches; (b) comparison of the x-axis distribution of good matches, showing the level of uniform distribution.

There are two cases that illustrate the feature matching comparison between SIFT and LoFTR in Figure 7, and the distribution of red dots, indicative of matching points between adjacent images. Figure 7a and 7b illustrate the SIFT-based feature matching technique, where the red dots are sparser and primarily concentrated within the soil pixels. Conversely, the

LoFTR-based feature matching (Figure 7c and 7d), demonstrates a dense and more evenly spread distribution of red dots, extending well into the areas covered by the plants. This uniformity suggests a superior detection of matching features across both the soil and the vegetation, contributing to a more accurate homograph estimation that is crucial for seamless image stitching. For the purpose of compiling multiple images to count pods on a plot scale, the effectiveness of LoFTR is clear, as it provides a stronger foundation for constructing a continuous and precise representation of the field, which is essential for accurate pod enumeration and subsequent agricultural assessments.

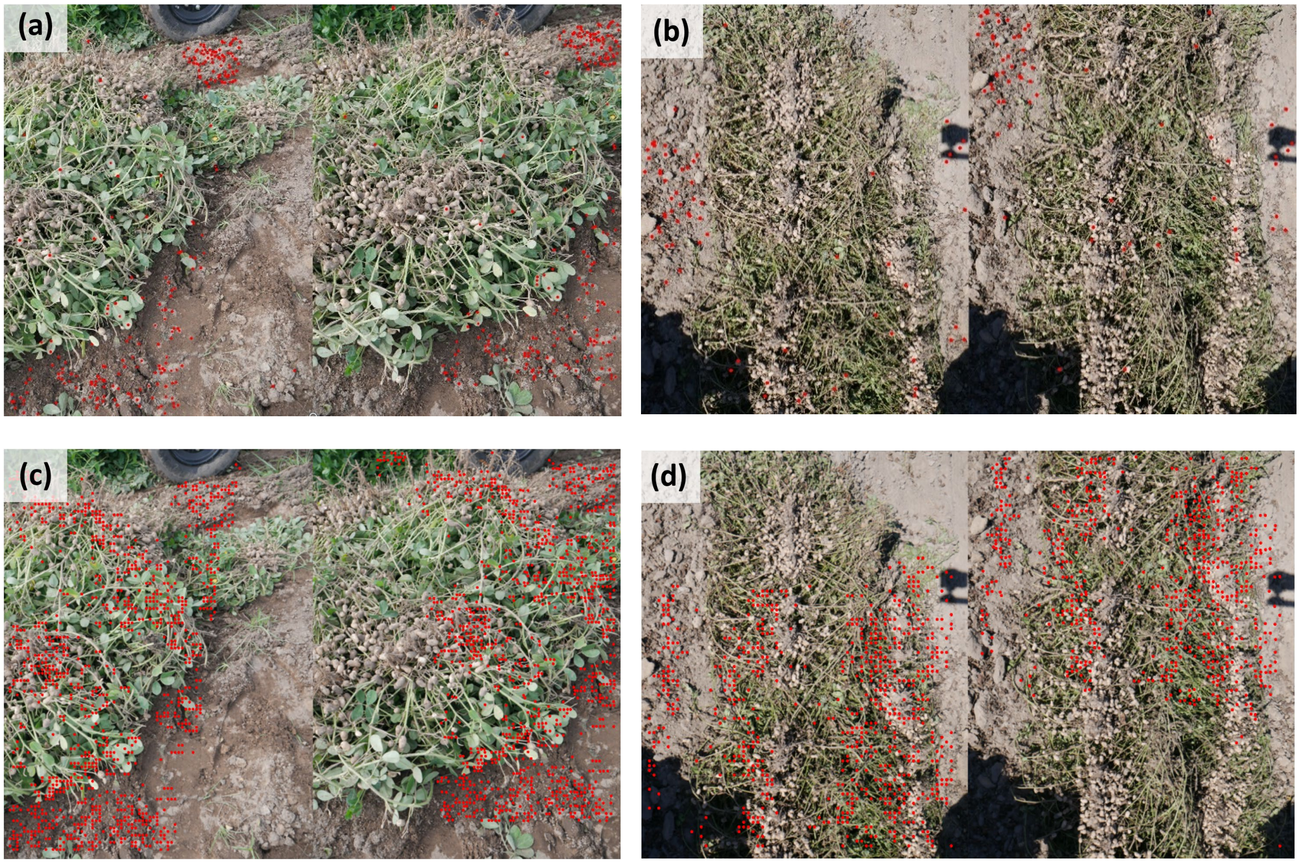


Figure 7. Comparison of feature matching between the SIFT and LoFTR algorithm. (a) and (c) are generated from sift-based algorithm; (c) and (d) are from LoFTR-based algorithm. The Red dots are the matching points between the two adjacent images.

3.3 performance of Plot-scale image stitching

In order to evaluate the performance of proposed Hierarchical stitching strategies for plot-scale image stitching algorithm, we compared the different configuration of strategies (sequences stitching and Hierarchical stitching) and matching algorithms (SIFT and LoFTR). Figure 8 showcases a comparison of different image stitching methods applied to a sequence of agricultural field images. Figure 8a and 8b demonstrate sequence stitching using SIFT and LoFTR respectively, illustrating how each method performs in aligning and merging multiple images into a single, wider panoramic view. While Figure 8c and 8d present the results of hierarchical stitching strategy, again comparing SIFT and LoFTR methods. The objective here is to evaluate the effectiveness of the stitching algorithms in terms of seamlessness, accuracy, and the ability to handle potential distortions and overlaps in the series of images.

Compared to the simpler sequential methods (a) and (b), which stitch images using SIFT and LoFTR in a sequence fashion, the hierarchical approach with LoFTR presents a marked improvement. The hierarchical strategy can avoid the cumulative perspective transform of stitching images sequentially. Sequential methods may introduce imperfections such as misalignments and artifacts, compromising the accuracy of subsequent analyses. Method (c) represents a step up by integrating the hierarchical strategy with SIFT, but still doesn't quite match the finesse of method (d). LoFTR's transformer-based deep learning architecture excels in detecting complex patterns in agricultural fields, which are rich with subtle yet vital indicators of crop health and development. In summary, method (d) not only delivers the most coherent and visually accurate panorama but also stands out as a reliable asset for agronomists and farmers, potentially enhancing crop analysis and informing better decision-making.

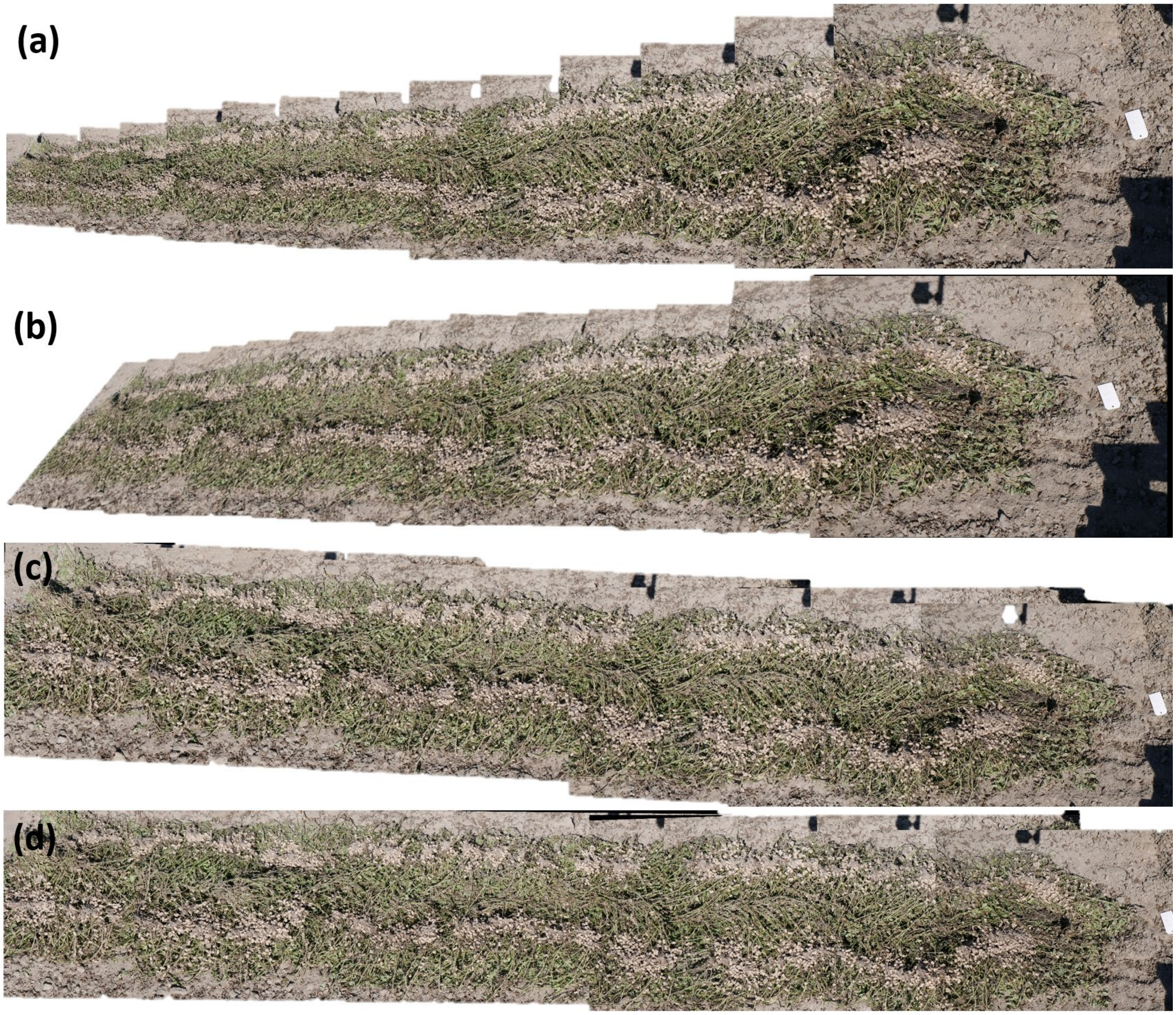


Figure 8. Comparison of different Image stitching method. (a) sequence stitching with SIFT; (b) tsequence stitching with LoFTR; (c) hierarchical strategy with SIFT; (d) hierarchical strategy with LoFTR.

3.3 Peanut Pod Detection evaluation

3.3.1 Performance matrics

Performance of the Peanut Pod detector was evaluated with COCO metrics (Everingham et al., 2010), as well as the Parameters (Params in millions) and its computational requirements (GFLOPs, Giga Floating Point Operations Per Inference). Specifically, the COCO metrics, including Precision, Recall, mAP50, and mAP95, were used and defined in the following equations:

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

$$IoU = \frac{area\ of\ overlap}{area\ of\ union} \tag{10}$$

$$AP = \sum_{n=1}^{N} P(n)\Delta r(n) \tag{11}$$

where *TP*, FP, and FN are the numbers of true positive cases, false positive cases, and false negative cases, respectively; *N* is the number of all test images; *P*(*n*) is the precision of *n* images and Δ*r*(*n*) is the change of recall from *n*-1 to *n*. In the COCO metrics, the standard IoU threshold used to define whether a detection is a true positive is 0.5. For mAP50 and mAP95, the calculation is an average of the Average Precision (AP) for different classes at a specific IoU threshold (50% for mAP50 and 95% for mAP95).

3.3.2 Performance comparation among other models

In order to verify the performance of our customized RT-DETR model, we compared its peanut pod detection accuracy against the state-of-the-art object detection models, especially YOLOv8 (*ultralytics*), and RT-DETR with different backbones (Lv et al., 2023) and YOLO-DETR (Ouyang, 2024). We consider these lightweight models for in-field application.

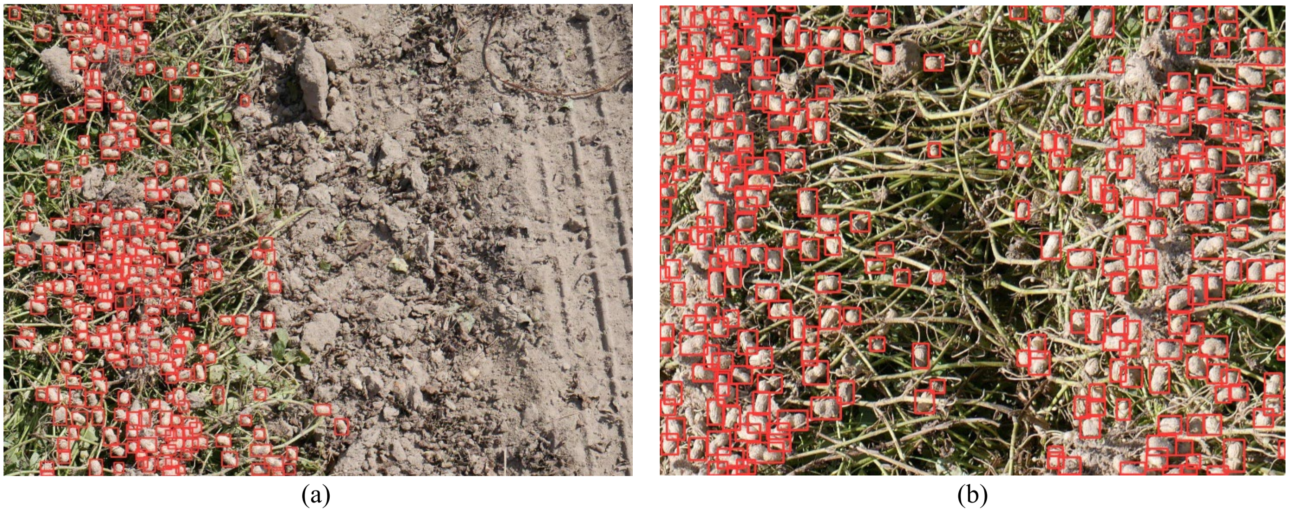
The customized model stands out with the highest scores in precision (86.9%), recall (85.1%), mAP50 (89.3%), and mAP95 (55.0%), indicating superior detection accuracy for peanut pods at both lenient and strict evaluation thresholds (Table 3). Remarkably, this high level of performance is achieved with fewer parameters (16 million) and a lower computational cost (48.1 GFLOPs) relative to some other top-performing models. The YOLOv8 models show a range of performances with a trade-off between accuracy and efficiency, while the DETR-augmented variants exhibit a notable reduction in precision and recall compared to their counterparts. Overall, the custom model offers an optimal balance between detection accuracy and computational efficiency, making it potentially the most practical choice for real-world applications where resources may be limited.

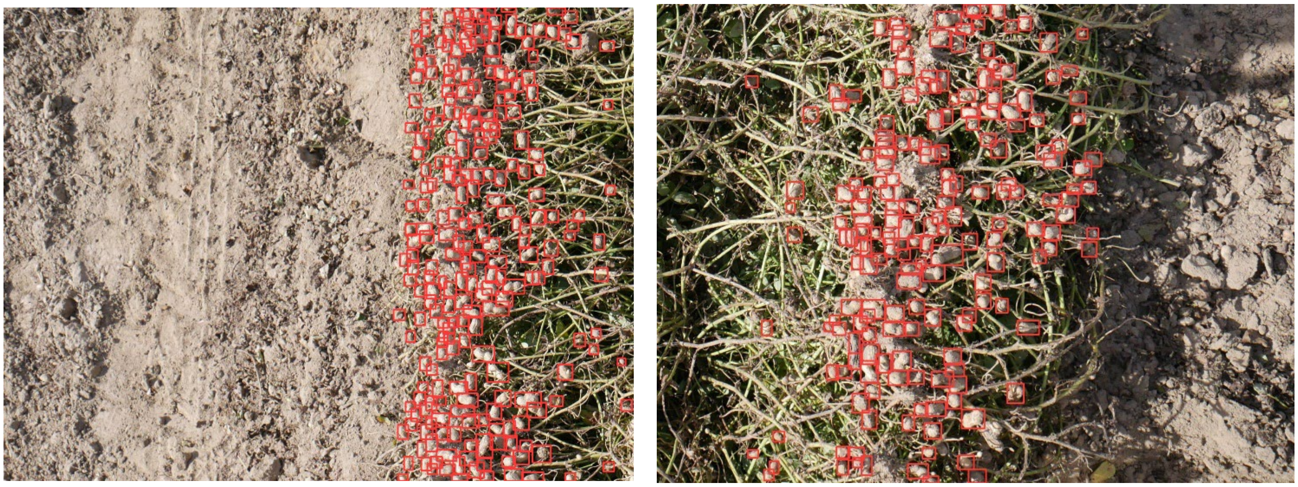
Table 3. Comparative performance of recent lightweight object detectors and our customized models

Model	Precision	Recall	mAP50	mAP95	Params (M)	GFLOPs
YOLOv8n	81.7	80.3	84.3	45.0	3.0	**8.1**
YOLOv8s	86.4	83.9	88.5	50.4	11.1	28.4
YOLOv8n-DETR	79.6	74.7	80.5	42.6	**6.1**	11.7
YOLOv8s-DETR	84.4	80.9	85.5	48.3	12.9	27.2
RTDETR-l	82.5	78.5	84.4	45.9	32.0	103.4
RT-DETR-x	86.3	83.2	87.6	52.0	65.4	222.5
RT-DETR-r18 (baseline)	84.5	80.6	86.0	49.1	19.9	56.9
Ours	**86.9**	**85.1**	**89.3**	**55.0**	16.0	48.1

The effectiveness of a peanut pod detection algorithm across various agricultural scenarios were shown in Figure 9, with red rectangles highlighting the detected pods in images (a) through (d). The algorithm demonstrates an impressive ability to adapt to different environmental conditions, from sparse to lush foliage, highlighting its robustness in detecting pods of a certain size range. This adaptability is key for potential applications in automating the quantification of peanut pod yields.

Despite the algorithm's proficiency, the illustration also sheds light on the challenges it faces. Difficulties primarily occur when pods are occluded by other pods, leaves, or stems, which may lead to incomplete or inaccurate detections. Factors such as heavy occlusion, shadows, low contrast environments where pods do not stand out from the soil, and the presence of stones that mimic the shape and size of pods can lead to misidentification. These limitations highlight the need for further refinement of the algorithm to ensure reliable pod detection amidst the complexities of real-world agricultural settings.


(a) (b)



(c) (d)

Figure 9: Illustration of peanut pod detection. The red rectangle is the predicted pods in the images.

3.3.2 Ablation Studies

We performed ablation experiments on our customized RT-DETR (Table 4), which determined the necessity of optimized modules in the learning architecture. In summary, each module brings distinct advantages to the table: FasterBlock minimizes the computation and provides better feature map, DySample and ADown module improve fusion of the multi-scale features.

Specifically, the inclusion of FasterBlock demonstrates moderate improvements in precision and recall while reducing the model's complexity and computational cost. DySample's implementation notably increases precision and recall and provides a small boost in mAP50, albeit at the cost of higher computational load. The addition of ADown yields substantial enhancements across all performance metrics, especially mAP50 and mAP95, without a significant increase in computational demands. Remarkably, the integrated application of all three components results in the best mAP95 outcome, underscoring their synergistic effect on the model's stringent precision. Moreover, this amalgamation leads to a leaner model with the least computational requirements, highlighting the efficacy of the combined modifications in optimizing the RT-DETR for cluster detection tasks.

Table 4. Ablation studies of customized RT-DETR

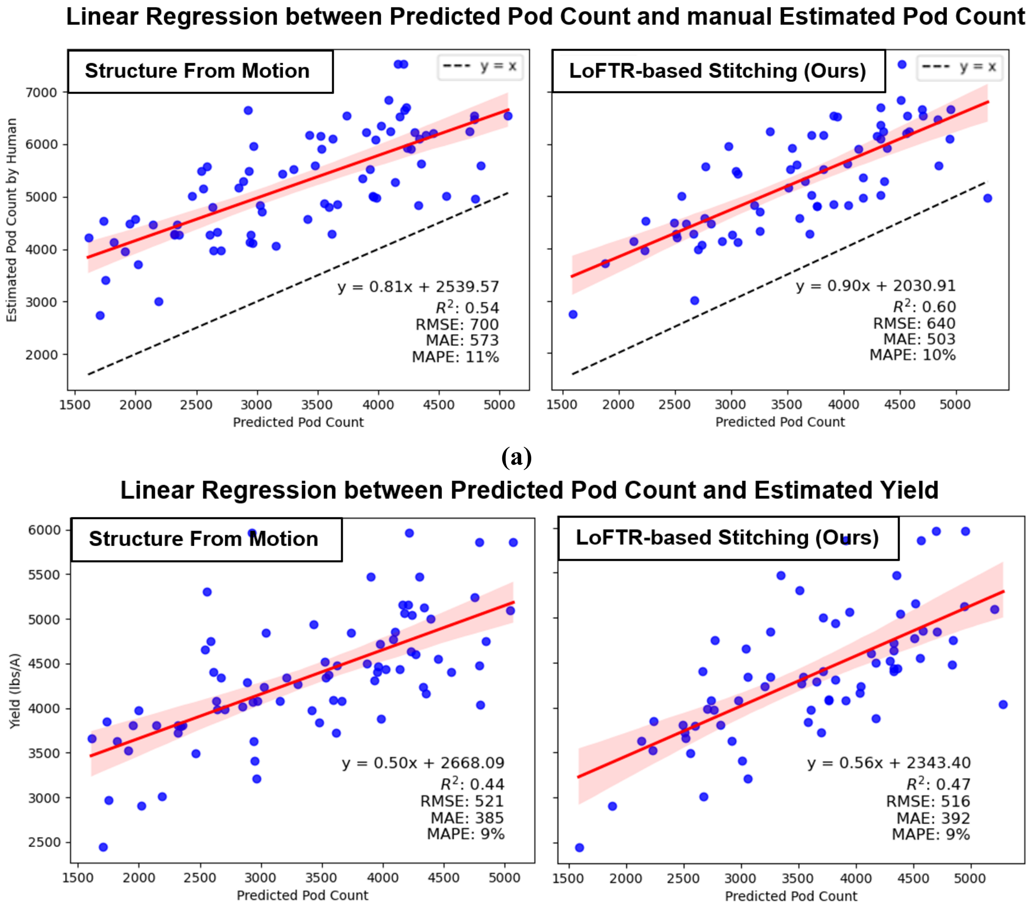
FasterBlock	DySample	ADown	Precision	Recall	mAP50	mAP95	Params (M)	GFLOPs
×	×	×	84.5	80.6	86.0	49.1	19.9	56.9
✓	×	×	85.5	83.8	88.0	53.2	16.8	49.8
×	✓	×	86.5	84.7	88.7	52.4	19.8	57.2
×	×	✓	**87.5**	85	**89.6**	53.9	19.7	57.3
✓	S✓	✓	86.9	**85.1**	89.3	**55.0**	**16.0**	**48.1**

3.4 Yield estimation

To assess the capability of our proposed pipeline, we conducted a manual estimated process on 82 plots with distinct genotypes. These individual plot-scale images were stitched and detected with the improved RT-DETR model. Another useful offline stitching approach based on structure using Agrisoft Metashape from motion approach was compared.

Structure from Motion (SfM) with Agrisoft Metashape is a process where a series of overlapping images, usually captured from a UAV or other platform, are algorithmically combined to reconstruct a 3D scene in the form of a textured mesh. Metashape detects distinctive features in these images and matches them to determine the camera's position and orientation at the time of capture, resulting in a detailed 3D point cloud. This cloud is then transformed into a mesh and overlaid with the original imagery to produce a high-resolution orthomosaic, a geographically corrected, uniform-scale image that accurately depicts the terrain and can be utilized for applications like precision agriculture and land management. However, it requires huge computations and processing time.

The linear regression analyses conducted in this study illuminate the relationship between pod counts estimated through Structure-from-Motion (SFM) techniques and our proposed method both in actual pod counts and yield measurements. The use of the LoFTR-based stitching method demonstrates an improvement in the correlation between predicted pod counts and actual yield when compared to the Structure from Motion technique. This advancement is evident in the tighter confidence intervals and a more coherent trendline in the LoFTR-based approach, which suggests a more reliable prediction of yield from the pod counts obtained. This could signify a substantial step forward in precision agriculture, particularly in automating and enhancing the accuracy of yield predictions based on pod counts. For example, the average weight of unit pods among different genotypes. These results underscore the necessity of incorporating a broader range of data, including detailed phenotypic traits and environmental conditions, to enhance the accuracy of yield forecasts in precision agriculture.



(a)

(b)

Figure 10: Linear Regression of pod number and yield using the predicted plot-level pod counts. (a) comparison of pod number regression using the predicted pod number of peanut plot between SfM-based orthomosaic and LoFTR-based plot-scale stitching images. (b) comparison of yield regression using the predicted pod number of peanut plot between the two methods.

4. Conclusions

In conclusion, our research marks a significant advancement in the field of precision agriculture and peanut breeding. By integrating a mobile robotic imaging system with cutting-edge image processing algorithms, we have successfully automated the yield determination process in plot-scale peanut crops. The hierarchical plot-scale image stitching method, underpinned by the Local Feature Transformer (LoFTR), outperformed traditional SIFT-based approaches in stitching accuracy. Furthermore, the refined Real-Time Detection Transformer (RT-DETR) model demonstrated superior detection capabilities, with a notable increase in pod detection accuracy with a smaller computational expense compared to the baseline. When applied to pod count and yield prediction, our system achieved a regression MAPE that surpassed the Structure from Motion (SFM)-based approaches, highlighting its robustness and effectiveness. This technological leap not only streamlines the yield determination process but also presents an invaluable tool for breeders to efficiently select high-yield genotypes. The ability of our system to accurately quantify yield in complex scenarios underscores its potential to transform peanut breeding and contribute significantly to global food security.

Acknowledgements

This study is supported by the University of Florida IFAS LIFT AI seed grant.
The authors gratefully thank Nolan Gao and Tiger Cao for labeling the peanut pod dataset.

References

Akyon, F. C., Altinuc, S. O., & Temizel, A. (2022). Slicing aided hyper inference and fine-tuning for small object detection. 2022 IEEE International Conference on Image Processing (ICIP),

Bagherian, K., Bidese-Puhl, R., Bao, Y., Zhang, Q., Sanz-Saez, A., Dang, P. M., Lamb, M. C., & Chen, C. (2023). Phenotyping agronomic and physiological traits in peanut under mid-season drought stress using UAV-based hyperspectral imaging and machine learning. *The Plant Phenome Journal*, *6*(1), e20081.

Balota, M., & Oakes, J. (2016). Exploratory use of a UAV platform for variety selection in peanut. Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping,

Chen, J., Kao, S.-h., He, H., Zhuo, W., Wen, S., Lee, C.-H., & Chan, S.-H. G. (2023). Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Dobreva, I. D., Ruiz-Guzman, H. A., Barrios-Perez, I., Adams, T., Teare, B. L., Payton, P., Everett, M. E., Burow, M. D., & Hays, D. B. (2021). Thresholding analysis and feature extraction from 3D ground penetrating radar data for noninvasive assessment of peanut yield. *Remote Sensing*, *13*(10), 1896.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*, 303-338.

Larsen, J. C., Austin, R., Dunne, J., & Kudenov, M. W. (2022). Drone-based polarization imaging for phenotyping peanut in response to leaf spot disease. Polarization: Measurement, Analysis, and Remote Sensing XV,

Liu, W., Lu, H., Fu, H., & Cao, Z. (2023). Learning to Upsample by Learning to Sample. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Lv, W., Xu, S., Zhao, Y., Wang, G., Wei, J., Cui, C., Du, Y., Dang, Q., & Liu, Y. (2023). Detrs beat yolos on real-time object detection. *arXiv preprint arXiv:2304.08069*.

Manley, A., Ravelombola, W., Cason, J., Bennett, B., Pham, H., Kimura, E., Ruhl, C., Ahmad, W., & Brown, M. (2023). Use of Unmanned Aerial System (UAS) Phenotyping to Predict Pod and Seed Yield in Organic Peanuts. *American Journal of Plant Sciences*, *14*(3), 415-426.

Ouyang, H. (2024). DEYO: DETR with YOLO for End-to-End Object Detection. *arXiv preprint arXiv:2402.16370*.

Patrick, A., Pelham, S., Culbreath, A., Holbrook, C. C., De Godoy, I. J., & Li, C. (2017). High throughput phenotyping of tomato spot wilt disease in peanuts using unmanned aerial systems and multispectral imaging. *Ieee Instrumentation & Measurement Magazine*, *20*(3), 4-12.

Puhl, R. B., Bao, Y., Sanz-Saez, A., & Chen, C. (2021). Infield peanut pod counting using deep neural networks for yield estimation. 2021 ASABE Annual International Virtual Meeting,

Sarkar, S., Cazenave, A.-B., Oakes, J., McCall, D., Thomason, W., Abbott, L., & Balota, M. (2021). Aerial high-throughput phenotyping of peanut leaf area index and lateral growth. *Scientific Reports*, *11*(1), 21661.

Sarkar, S., Cazenave, A. B., Oakes, J., McCall, D., Thomason, W., Abbot, L., & Balota, M. (2020). High-throughput measurement of peanut canopy height using digital surface models. *The Plant Phenome Journal*, *3*(1), e20003.

Sarkar, S., Ramsey, A. F., Cazenave, A.-B., & Balota, M. (2021). Peanut leaf wilting estimation from RGB color indices and logistic models. *Frontiers in Plant Science*, *12*, 658621.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Proceedings of the IEEE conference on computer vision and pattern recognition,

Sun, J., Shen, Z., Wang, Y., Bao, H., & Zhou, X. (2021). LoFTR: Detector-free local feature matching with transformers. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Tellus, U. (2021). *At ARS, Peanut Research is Alive and Well*. Retrieved March 15 from https://tellus.ars.usda.gov/stories/articles/ars-peanut-research-alive-and-well

ultralytics.

Wang, C.-Y., Yeh, I.-H., & Liao, H.-Y. M. (2024). YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv preprint arXiv:2402.13616*.

Xu, R., & Li, C. Y. (2022). A modular agricultural robotic system (MARS) for precision farming: Concept and implementation [Article; Early Access]. *Journal of Field Robotics*, 23. https://doi.org/10.1002/rob.22056

Yuan, H., Bennett, R. S., Wang, N., & Chamberlin, K. D. (2019). Development of a peanut canopy measurement system using a ground-based lidar sensor. *Frontiers in Plant Science*, *10*, 203.

Yuan, H., Wang, N., Bennett, R., Burditt, D., Cannon, A., & Chamberlin, K. (2018). Development of a ground-based peanut canopy phenotyping system. *IFAC-PapersOnLine*, *51*(17), 162-165.