

2
3 **Aerial-Based Generalized Plant Counting Algorithm for Efficient Crop Emergence Estimation**

4
5 **Ramsha Shahid¹, Waqar S. Qureshi^{1,3}, Umar S. Khan^{1,2}, Arslan Munir⁴, Ayesha Zeb¹, Syed Imran**
6 **Moazzam^{1,2}**

7 ¹Department of Mechatronics Engineering, National University of Sciences and Technology, H-12,
8 Islamabad, Pakistan

9 ²Robot Design and Development Lab, National Centre of Robotics and Automation (NCRA), National
10 University of Sciences and Technology, H-12, Islamabad, Pakistan

11 ³School of Computer Science, Technological University Dublin, Dublin, Ireland

12 ⁴Kansas State University, Manhattan, Kansas, USA

13 Corresponding author: Waqar S. Qureshi (e-mail: waqar.qureshi@tudublin.ie).

14
15 **ABSTRACT**

16 Crop emergence estimation at early crop growth stages is becoming increasingly important for the long-term
17 sustainability of natural resources. It helps farmers and agricultural stakeholders in the efficient allocation of
18 resources like water, pesticides, and fertilizers. It can be used to estimate the yield and seed quality, identify
19 the region of potential yield losses, and make future agriculture plans. These future agriculture plans can play
20 a crucial role in ensuring maximum crop population and yield while utilizing the same limited land and natural
21 resources. Most of the existing plant counting frameworks require offline processing of images with licensed
22 software to develop orthomosaic using multiview stereo and are computationally expensive. This study
23 proposed a generalized plant counting framework capable of onboard processing and directly estimates counts
24 from aerial images. It consists of three core modules: overlap detection, plant detection, and plant counting.
25 The overlap detection module replaces the need for computationally expensive orthomosaic formation to
26 avoid counting repetition by overlap masking based on only visual cues. Three different methods are
27 evaluated as core modules for finding an optimal generalized solution for plant counting based on time
28 complexity and accuracy. In the first method after overlap detection semantic segmentation with U-NET is
29 employed as plant detection modules by counting connected pixels. Object detection with YOLOv7 is utilized
30 as a plant detection module after overlap removal in the second method. The third method involves a counting
31 framework based on multiple object tracking using YOLOv7 for object detection and SORT for object
32 tracking as a replacement for the overlap detection module that has the potential for a real time applicable
33 system. The proposed algorithm is evaluated on two distinct tobacco fields. The high-resolution aerial data is
34 collected from tobacco fields near Peshawar, Pakistan, and is human labeled. First and second methods show
35 average F1 scores of 0.947 and 0.9667 respectively whereas the third method has the potential for real-time
36 applicability with an average F1 score of 0.967.

KEYWORDS semantic segmentation, plant count, deep learning, U-Net, overlap detection, object tracking, object detection, YOLO, SORT

1. INTRODUCTION

Crop emergence estimation at early stages of plant growth can play an important role in the optimization of agriculture farming. In the field of agriculture, the optimal utilization of limited natural resources such as water and land is crucial to meet the growing demand of our world. However, the rapid population growth ranging from 7.7 billion in 2019 to 10.9 billion in 2100 not only causes environmental deterioration but also poses a significant threat to the long-term sustainability of the natural resources (Maja & Ayano, 2021). To achieve the optimal utilization of limited natural resources and maximize the yield, it becomes essential to ensure that every seed planted emerges successfully and contributes to the overall yield.

Crop emergence estimation at early growth stages enables timely interventions, such as replanting or applying proper agricultural input at the affected zone. By taking proactive measures based on early crop emergence estimation, potential yield losses can be mitigated, leading to improved agricultural productivity and resource utilization. Plant counting by traditional methods is a time-consuming, labor-intensive process and prone to human error. Therefore, there is a crucial need for automation to enhance the effectiveness of plant counting. Implementation of automated plant counting systems has the potential to streamline agricultural operations, improve efficiency, and enable farmers to make informed decisions based on accurate and reliable plant count data. Advanced fields like deep learning (DL) or machine learning with computer vision can be utilized to automate plant counting and generate accurate results on time (Kamilaris & Prenafeta-Boldú, 2018).

Extensive field monitoring is necessary to conduct crop stand count analysis, which can only be accomplished via Sentinel data (the Sentinels, which are satellites operated by the European Space Agency (ESA), have been specifically developed to provide an extensive range of data and imagery as part of Europe's Copernicus program) or unmanned aerial vehicle (UAV) imagery (Mancini et al., 2019). In comparison to UAV data, sentinel data requires additional preprocessing, such as cloud calibration and polygon construction, as well as greater storage capacity. Data from UAVs can be acquired at any favorable time and height. The utilization of UAV imagery in agriculture has become increasingly popular in recent years (Ghazali et al., 2022). Researchers have been exploring the combination of UAV imagery with advanced algorithms (Jha et al., 2019; Li et al., 2020; Saleem et al., 2021) such as deep learning (DL), image processing, object detection, and machine learning algorithms to automate various agricultural processes such as crop monitoring, disease detection, yield estimation, pest management, and plant counting.

A deep learning based framework for plant counting using UAV imagery is proposed in this paper. Although the framework is specifically tested on tobacco fields, its applicability extends to other crop types. Pakistan's tobacco industry comprises a substantial network of 50,000 growers distributed throughout the nation, facilitating the export of significant quantities of diverse tobacco types and products. This tobacco sector is also one of the major contributors to the national exchequer which contributed more than Rs.124 billion in Federal Excise Duty/ Sales Tax in FY 2019-2020 (*Potential for Export | PTB*, n.d.). Additionally, apart from its economic value, tobacco is also recognized for its therapeutic properties (Charlton, 2004). Crop emergence estimation involves the sequential process of plant detection and counting. Previous studies have utilized efficient techniques such as image processing, machine learning, and DL-based methods to achieve accurate results.

Image processing-based contour detection method (Rahmawati et al., 2021) followed by morphological operations is proposed for Tobacco counting. The tobacco detection algorithms (Fan et al., 2018) involve extracting tobacco plant regions in an image using morphological operations and watershed segmentation, followed by training a DL model to classify these regions into tobacco and non-tobacco plant regions. (Y. Wang et al., 2022) evaluated EXG (Excess Green Index), NGRDI(Normalized Green-Red Difference Vegetation Index), and EXG-EXR (Excess Green Minus Excess Red Index) for optimal segmentation threshold of the Tobacco plant in an orthophoto. (Liu et al., 2016) used the Otsu and EXG followed by morphological operation for wheat counting. (Shirzadifar et al., 2020) after mosaicking images, performed segmentation with two different techniques EXG method and k-mean clustering for stand count estimation of maize crop. These image-processing-based techniques involve thresholding and contour detection; it is extremely difficult to find an optimal value for a threshold and contour for vegetation other than crops, such as weeds and grass, resulting in extensive post-processing after detection. (Oh et al., 2020) utilized the YOLOv3 object detection algorithm to assess cotton growth, while (Pang et al., 2020) developed a system for early-season maize stand count detection using Max Area Mask Scoring RCNN and segmentation. (Miao et al., 2020) used DL-based semantic segmentation for the segmentation of sorghum plants by classifying every pixel of hyperspectral images to the organ level. Semantic segmentation of RGB images using U-Net architecture was utilized by (Nee et al., 2021; Kitano et al., 2019).

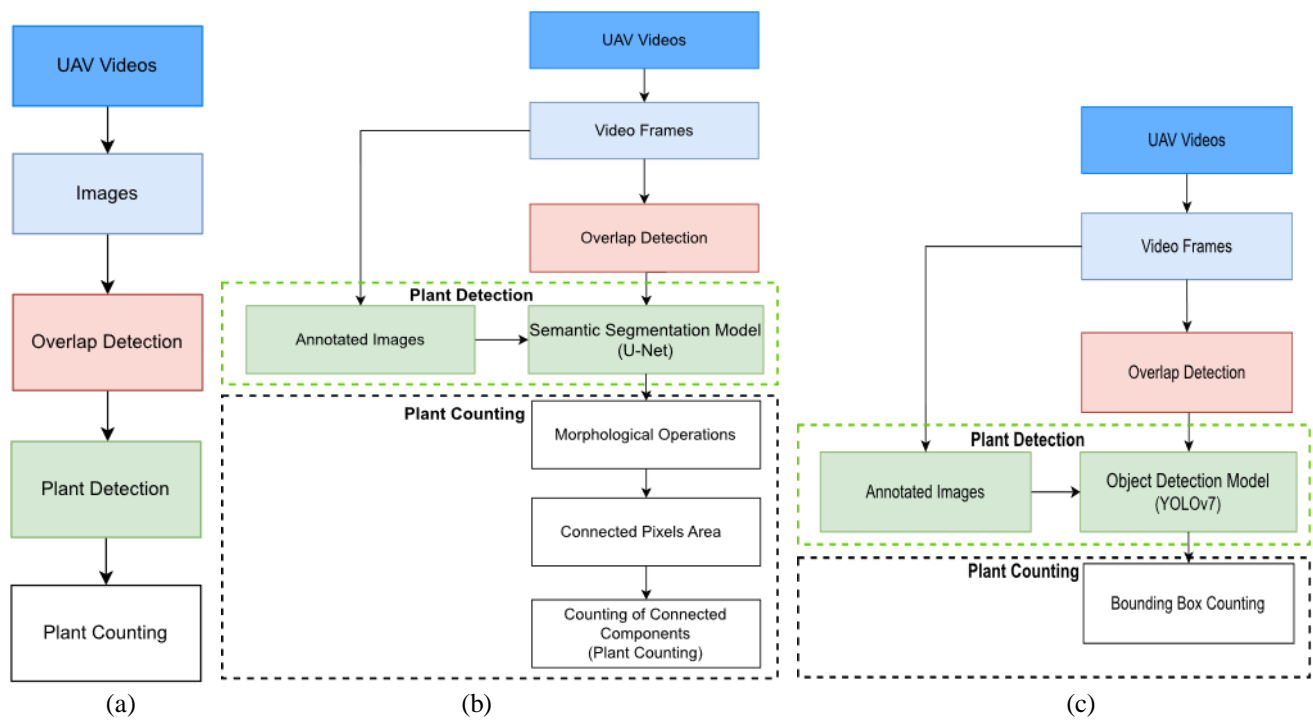


Fig. 1. Proposed general framework with different DL architectures for crop emergence estimation. (a) proposed general framework consisting of six modules, (b) proposed framework with semantic segmentation for plant detection, (c) proposed framework with object detection for plant detection. The proposed framework is represented by distinct block colors, with each color representing a specific module. The dashed green block denotes all the sub-modules within the plant detection module, while the dashed black block represents the sub-modules within the plant counting module.

(Valente et al., 2020) utilized Otsu thresholding for segmentation of spinach plants after conversion of orthomosaic to smaller units and AlexNet for assessing the number of pixels per plant. A two-branched CNN based architecture was evaluated by

(Osco et al., 2021) for a per image plant count for orchard and corn fields. In DL-based approaches, object detection and semantic segmentation show promising results. However, all these techniques typically require a computationally intensive process of orthomosaic formation. Commercialized image processing software is used for this orthomosaic formation. Orthomosaic or orthophoto generation involves camera pose estimation and multiview stereo reconstruction, making the system inapplicable in real time. In addition, several research are found in literature that have employed multi object tracking techniques for fruit/plant counting to develop a real-time system.

(Parico & Ahamed, 2021), utilized YOLOv4 in combination with DeepSORT for pear counting and tested on a mobile phone video having 2-3 pears and faced illumination challenges. (Tan et al., 2022) utilized YOLOv4 with optical flow for counting cotton seedlings on a dataset recorded by handheld system. (Egi et al., 2022) applied the integration of YOLOv5 and DeepSORT for tomato counting in a developed experimental field. Furthermore, (Yang et al., 2022) evaluated the combination of CenterNet and DeepSORT on a dataset captured from a mobile phone, which included only one-row crop. It has been observed that most of the proposed framework were evaluated on a dataset collected from a handheld systems/mobile phone that limits its feasibility for large-scale field monitoring. In the study by (Feng et al., 2020), a preprocessing pipeline was developed to geo-reference crop rows, eliminating the requirement for orthomosaic formation.

The necessity of orthomosaic formation in state-of-the-art research typically involves high computational power and commercial software. This process of orthomosaic formation typically involves offline processing, which diminishes the benefits of real-time plant counting. With promising results of using UAV imagery with advanced fields, this study aims to develop a generalized plant counting framework. The proposed framework consists of a novel overlap detection module based on visual cues only that omits the need for a computationally expensive orthomosaic formation process. The feasibility of semantic segmentation and object detection in combination with the overlap detection module for tobacco plant count estimation is assessed. The results generated in lesser time enhance the benefits of plant counting such as appropriate distribution of resources at the early stage of crop growth, and enabling farmers to make informed decisions regarding irrigation, fertilization, and pest management. A real-time applicable method is also proposed by integrating the object detection algorithm with SORT tracking. A real-time plant counting framework can facilitate the early detection of plant diseases, pests, or other issues to prevent potential crop losses.

2. MATERIALS AND METHOD

The proposed general framework for crop emergence estimation is depicted in **Fig. 1a**. The framework comprises six modules: input video, frame extraction, overlap detection, plant detection, and counting. The frames are first extracted from the video stream and then resized to reduce processing time. Second, all of these frames are subjected to overlap detection, which masks the overlapping region between consecutive frames. The framework consists of a pipelined stage of DL network architectures for plant detection. We have evaluated different architectures to find the best block for the framework. Finally, the counting of these detected plants

2.1 Data Collection

We have acquired a new Tobacco field dataset. Aerial data is recorded from two distinct tobacco fields near Peshawar, Pakistan (34.332055° N, 71.95546° E) with 20-30 frames per second using a DJI Mavic Mini drone which is equipped with a high-resolution onboard RGB camera. The tobacco plants are captured at 1920 x 1080 resolution at the early growth stage of 15-40 approximately after emergence. The dataset is captured at an average altitude of 5 meters. **Fig. 2** shows images of a tobacco dataset in 1920 × 1080-pixel resolution under different soil textures and sunlight conditions. Two distinct tobacco fields are captured at

133 20-30 frames per second which results in two distinct field videos. These two videos are then split into smaller distinct clips for
134 ease of evaluation and ground truth analysis.



135 **Fig. 2.** Tobacco dataset images under different lighting and soil conditions

136 These clips are named as DATASET-1, DATASET-2, DATASET-3, DATASET-4 and DATASET-5 throughout this study. All
137 these datasets consist of one-directional movement of drone. Plant counting is most effective at this stage since production may
138 be boosted by correct agriculture input and weed infestation is relatively low at these early stages of growth. The speed and height
139 variations caused by the UAV's manual control result in an uneven overlap between the frames. Plant counting might be simple
140 if non-overlapping frames could be retrieved, as there would be no counting repetition due to overlap. An overlap detection
141 technique is designed to overcome this problem.

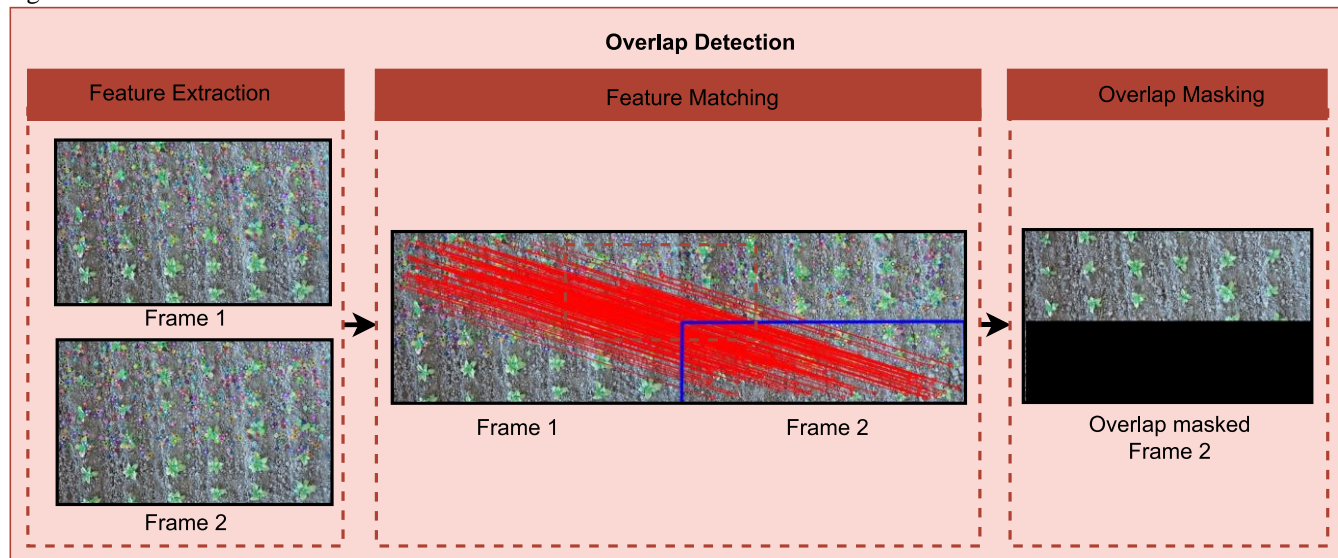
142 **2.2 Overlap Detection**

143 Frames are extracted from the video clips with 40-50% overlap between frames. The identification of overlapping regions between
144 consecutive frames is a crucial step in the process of plant counting. It is necessary to mark and determine these regions accurately
145 to ensure reliable counting results. By detecting and delineating the areas of overlap, subsequent counting processes can exclude
146 these regions, thus avoiding double-counting or miscounting of plants. To detect the region of overlap between consecutive
147 frames, the approach involves feature extraction, feature matching, and perspective transformation, as depicted in Figure 3. Each
148 step is discussed in the following sections.

149 **2.2.1 Feature Extraction**

150 Feature extraction has a major role in image processing. Features describe the relevant information which could be color,
151 shape, texture, size, etc. Features must contain all the important information to classify or identify objects/images (Kumar &
152 Bhatia, 2014). Features can have all the important information to represent any image thereby feature extraction could be called
153 a process of dimensionality reduction. In our study, feature extraction is utilized in the development of a methodology for
154 overlap detection as we must match the features between the frames for the identification of the overlap region. Many feature
155 extraction algorithms have been developed that extract features in the form of keypoints with descriptors. Keypoints are the
156 coordinates of the points having a fingerprint called a *descriptor* that defines the uniqueness of the transformation form of a
157 vector. Features might be scale, rotation, and illumination invariant. Some of these feature extractors are SIFT, SURF, KAZE,
158 ORB, etc. SIFT and SURF algorithms are not open-source, so we have used an open-source KAZE feature extractor
159 (Alcantarilla et al., 2012). It involves nonlinear diffusion filtering to obtain multiscale features that show much higher
160 repeatability and distinctiveness rates than SIFT and SURF algorithms that are based on the Gaussian scale space. As Gaussian
161 scale space smooths both noise and details at the same degree and does not respect the natural boundaries of the object, higher
162 distinctiveness allows for more reliable and accurate matching between features from different frames. By capturing unique
163 and discriminative characteristics of the image content, distinct features enhance the matching process, leading to more robust

164 and precise results. KAZE comparison with other feature extractors was done by (Tareen & Saleem, 2018) and their results
 165 have shown that it performs well in feature matching and is comparable to patented algorithms, such as SIFT, and SURF
 166 algorithms.



167 **Fig. 3.** Overlap detection and masking pipeline. Frame 1 and Frame 2 are two consecutive frames. Multicolor dots on frames in
 168 the feature extraction block represent some of the extracted features. The red lines in feature matching show the feature matching
 169 between frames, the blue rectangle highlights the identified overlapping region from homography, and the black solid rectangle
 170 in overlap masking block shows the masking of the identified overlap region.

171 **2.2.2 Feature Matching**

172 Matching the KAZE features of two consecutive frames is performed using a Flann-based matcher. It includes a set of
 173 algorithms that are optimized for fast nearest-neighbor searches in large datasets and high-dimensional features. It is faster
 174 than BFMATCHER for large datasets. The matches are sorted according to the distance and the first 500 matches are considered
 175 good matches. These good matches are then used for the estimation of homography.

176 **2.2.3 Homography**

177 The keypoints of sorted matches from the first frame/image are then considered as source points and keypoints of sorted
 178 matches from the second frame/image are considered as destination points and used for finding Homography between these
 179 frames. Homography relates the images of a plane taken by different orientations or positions of the camera. It is a 3-by-3
 180 matrix (H) in homogenous coordinates. Homography could be calculated for every match to find the solution with the least
 181 outliers. For the estimation of homography, RANSAC is used.

182 **2.2.4 RANSAC**

183 RANSAC (Random Sample Consensus) developed by (Fischler & Bolles, 1981) is a procedure for estimating the parameters
 184 of a mathematical model that is robust to noise/outliers. It can be fully described in three steps: (1) It samples the dataset into
 185 smaller datasets and considers these samples as the inliers, (2) Estimates the model using these samples, and (3) Calculates the
 186 score of inliers and outliers for the estimated model. These three steps are repeated to find the model with a greater number of
 187 inliers. The number of iterations depends on the probability of inliers, the probability of outliers, and the number of samples
 188 required for estimating a model. We have utilized RANSAC for the estimation of homography matrix.

189

2.2.5 Overlap Masking

Identification of overlap regions between consecutive frames is done by using the homography matrix, obtained through the above steps. The identified region is then masked from the image. The obtained results, as depicted in Figure 3, demonstrate the successful identification and masking of overlap regions from the images.

2.3 Plant Detection & Counting

The plant detection module is the core component of the framework. It leverages deep learning network architectures to detect plants in the frames. Multiple architectures are evaluated to identify the most suitable one for the framework's requirements. Tobacco plant accurate identification is highly important for the estimation of crop emergence. After the detection of these plants, a counting module is utilized to assess the number of identified tobacco plants.

2.3.1 Semantic Segmentation

Segmentation of images is one of the important techniques for different applications like object detection, autonomous vehicles, etc. Semantic segmentation is the segmentation of the whole scene in an image that is done by assigning labels to every pixel in an image. We have used LabelMe (Wada) for the annotation of images. The labeled dataset is created for two classes, viz. vegetation and non-vegetation containing 62 images for training, 22 images for validation, and testing on 13 images. We have utilized U-Net semantic segmentation architecture as detection module shown in **Fig. 1b**. The deep learning architecture U-Net (Ronneberger et al., 2015) is a type of convolutional neural network that was first used in biomedical applications for image segmentation. In a comparative study conducted by (Jeon et al., 2021), the U-Net outperformed SegNet, PSPNet, and DeepLab v3+ and other already available architectures for semantic segmentation. U-Net can extract features from low-resolution, and small-sized images and gives good results with smaller datasets. Due to these features, U-Net has become popular in agriculture domain. In a study by (Zhang et al., 2020), U-Net performed well in segmenting Purple Rapeseed leaves. (Moazzam et al., 2022) developed W-Net using two U-Nets that classify plants and weeds better than other segmentation models. Considering that our dataset solely consists of two classes, namely vegetation and ground, and does not include a distinct weed category, the W-Net architecture, which incorporates a second stage for weed and plant classification within the vegetation class is not the most suitable option for our specific application. In this context, it is more appropriate to opt for the U-Net architecture, which can effectively classify vegetation without the need for an additional stage. U-Net with VGG16 as its backbone is trained in this research using Google Colab. Semantic segmentation results in binary images by assigning one to vegetation pixels and zero to non-vegetation pixels. After segmentation of overlap region removed images, the resulting binary images are filtered using median blurring followed by morphological operation to remove noise from the predicted binary image. Counting is performed with connected pixel area which results in the total count of tobacco in images. To address the issue of partial plants across frames which potentially lead to inaccurate counts, a strategy is implemented. In this strategy, plants located at the boundaries of the frames are initially counted in one frame. However, in the following frame, the boundary plants are ignored to prevent double counting of partial plants and ensure more accurate counting results.

2.3.2 Object Detection

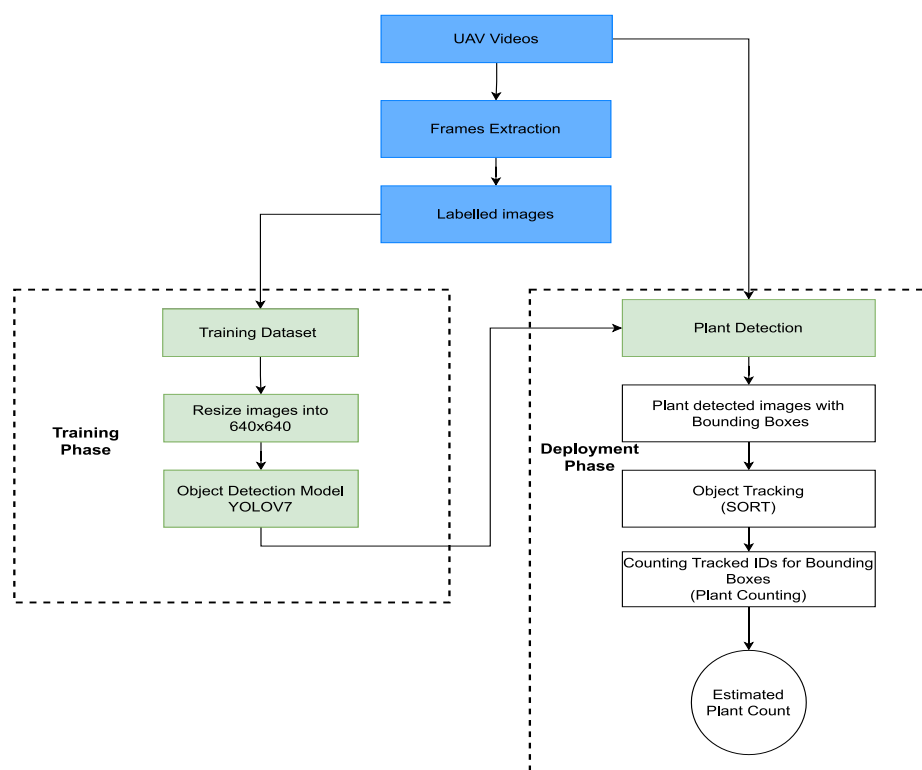
State-of-the-art object detection models are based on YOLO (You Look Only Once) (Redmon et al., 2016; Redmon & Farhadi, 2017; Bochkovskiy et al., 2020). We have evaluated YOLOv7 (C.-Y. Wang et al., 2022) for plant counting as shown in **Fig. 1c**. YOLOv7 is optimized with model reparameterization and dynamic label assignment without increasing the inference cost. In this paper, we have used pre-trained basic YOLOv7 rather than its scaled version. The model is trained on 83 images and images are annotated using LabelImg (Tzutalin, 2015)

228 After overlap detection, the frames are passed to trained YOLOv7, and then the number of bounding boxes are calculated to
 229 find the total count. The highest accuracy is achieved at a confidence threshold of 0.6. The problem of corner plants is much
 230 lesser than semantic segmentation as a bounding box is formed after correctly classifying something as tobacco rather than on
 231 connected pixels area.

232 The above-mentioned techniques based on semantic segmentation and object detection require very less time as compared to
 233 the state-of-the-art orthomosaic-based counting techniques. To make the system real-time we have evaluated a real-time plant
 234 counting framework using an object detection model with tracking for tobacco plant counting.

235 2.3.3 Object Detection & Tracking

236 We have also evaluated object detection combined with tracking shown in **Fig. 4** to make crop emergence estimation applicable
 237 in real-time. We have combined the trained YOLOv7 in the above section with the SORT(Simple Online and Real-time
 238 Tracking) (Bewley et al., n.d.). We have explored this multi-object tracking based framework for counting plants. This method
 239 does not involve overlap detection as detected plants are tracked which avoids the counting of the same plant multiple times
 240 due to overlap. We have used recorded videos for evaluation but this proposed framework has the potential to be utilized for a
 241 real-time system due to its online tracking system.



242 **Fig. 4.** Real-time plant counting framework using object detection and tracking.

243 2.4 Evaluation Metrics

244 For the semantic segmentation model intersection over union (IOU), Precision, and Recall are used as the evaluation metrics,
 245 **Eq. 1**, **Eq. 2** and **Eq. 3** show their mathematical representation, respectively. IOU is the most used metric for semantic
 246 segmentation. IOU values ranges from 0 to 1, where 0 represents the worst case that shows almost no overlap between ground
 247 truth and predicted object, and on the contrary, 1 represents 100 percent overlap between ground truth and prediction. Precision
 248 is the proportion of true positives (TP) to the total number of positive predictions (TP + FP). A high precision value indicates

accurate predictions with a low rate of false positives. Recall is the proportion of true positives (TP) to the total number of actual positive cases (TP + false negatives (FN)). A high recall indicates that the model predicts most positive cases as positive. TP, FP, and FN represent the number of correctly classified pixels of the target class, the number of pixels that are incorrectly classified as the target class, and mistakenly classified pixels of the target class as other classes or background, respectively.

$$IOU = \frac{\text{Area of Union}}{\text{Area of Overlap}} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

Precision, recall and F1 Score for plant counting are calculated using **Eq. 4**, **Eq. 5**, and **Eq. 6**. A precision score of 1 indicates that the counting method is highly accurate and reliable and that all identified plants can be assumed to be real. A recall score of 1 indicates that the counting method is highly sensitive and capable of detecting all of the plants present in the field whereas the F1 score of 1 indicates that all plants are accurately identified and counted (high recall), while also minimizing any false positives (high precision). F1 score of 1 signifies a perfect situation in which there are no overlooked plants or misidentifications made during the counting procedure.

$$\text{Precision} = \frac{\text{Number of correctly identified plants}}{(\text{Number of correctly identified plants} + \text{Number of incorrectly identified plants})} \quad (4)$$

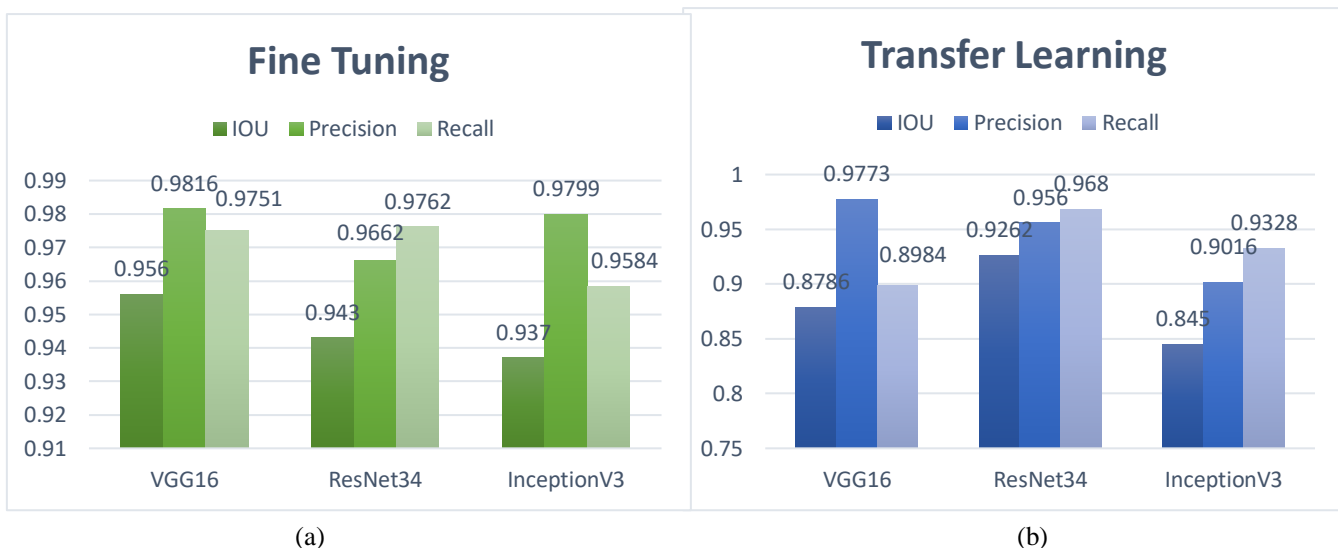
$$\text{Recall} = \frac{\text{Number of correctly identified plants}}{(\text{Number of correctly identified plants} + \text{Number of missed plants})} \quad (5)$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

3. RESULTS & DISCUSSION

3.1 Semantic Segmentation Evaluation

We have evaluated U-NET using **Eq. 1**, **Eq. 2** and **Eq. 3** for transfer learning (initializing the network with the pre-trained weights and freezing all layers except the fully connected layers), as well as fine-tuning (retraining all the layers of the network) as shown in **Fig. 5**. It can be observed clearly that for our problem we get better results from fine-tuning for RESNET34, VGG16, and Inception V3. As fine-tuning involves further training the pre-trained model results in a better accuracy with a smaller dataset thus making the model easily applicable to other plants too with lesser labeled data. U-Net with a fine-tuned VGG16 architecture as the backbone model demonstrates the highest IOU score of 0.9556. This indicates that boundaries of the plants in predictions closely align with the boundaries of the plants in the ground truths. Furthermore, this configuration also exhibits higher precision, reflecting the model's ability to minimize false positive predictions. Input image, U-NET prediction, and counting after morphological operations are shown for the initial/first frame with no overlap masked region in **Fig. 6a**, **Fig. 6b**, and **Fig. 6c**, respectively.



284

285

286

287

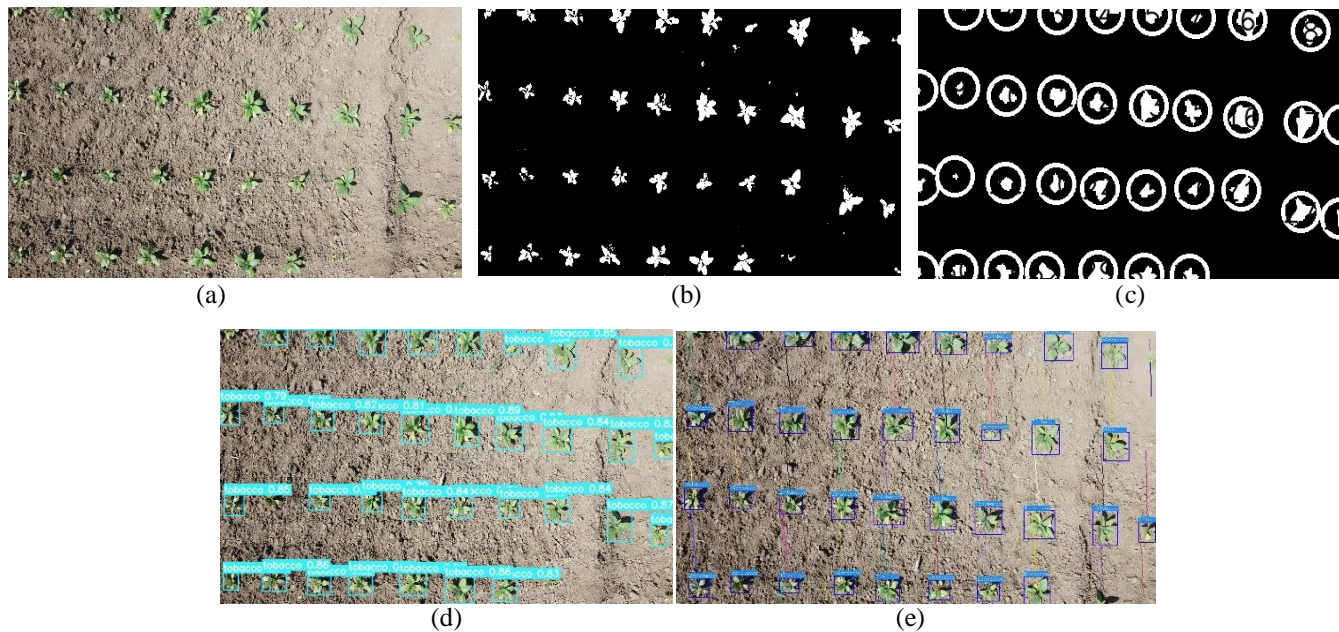
288

289

290

Fig. 5. Comparison of fine-tuning and transfer learning with different backbone models for the U-Net architecture: (a) Fine tuning, (b) Transfer learning.

The estimated plant count using semantic segmentation is shown in **Table I**. The model is evaluated on five distinct video clips of acquired three tobacco fields data. The distinct video clips are named Dataset 1, Dataset 2, Dataset 3, Dataset 4, and Dataset 5. The ground truth is the count done manually in a video, the estimated count is the results obtained from the framework, and the accuracy is the ratio of the estimated count to the ground truth.



291

292

293

294

295

296

297

298

299

300

301

302

303

Fig. 6. Visualization of results obtained using the proposed techniques on the same input image: (a) Input frame (first frame with no overlap) (b) Semantic segmentation prediction; (c) Counting results after segmentation; (d) Object detection on input frame; (e) Object detection with multiple object tracking.

304
 305
 306
 307
 308
 309
 310
 311
 312
 313
 314
 315
 316
 317
 318
 319
 320
 321
 322
 323
 324
 325
 326

Table I
Crop Emergence Estimation using Semantic Segmentation

Name	No. of images	True Count	Estimated	Precision	Recall	F1 Score
DATASET-1	28	472	428	1	0.906	0.951
DATASET-2	29	478	426	1	0.89	0.943
DATASET-3	32	522	510	1	0.977	0.988
DATASET-4	40	536	464	1	0.86	0.930
DATASET-5	61	540	463	1	0.857	0.923
Average	–	–	–	1	0.898	0.947

An average F1 score 0.947, a precision score of 1, and a recall score of 0.898 are observed overall for crop emerge estimation using semantic segmentation. It is observed that higher precision of the semantic model results in higher precision for counting which minimizes false positive detection, and a higher IOU of the U-Net will result in an accurate segmentation mask that reduces the likelihood of false positive and false negative plant detection leading to accurate counting.

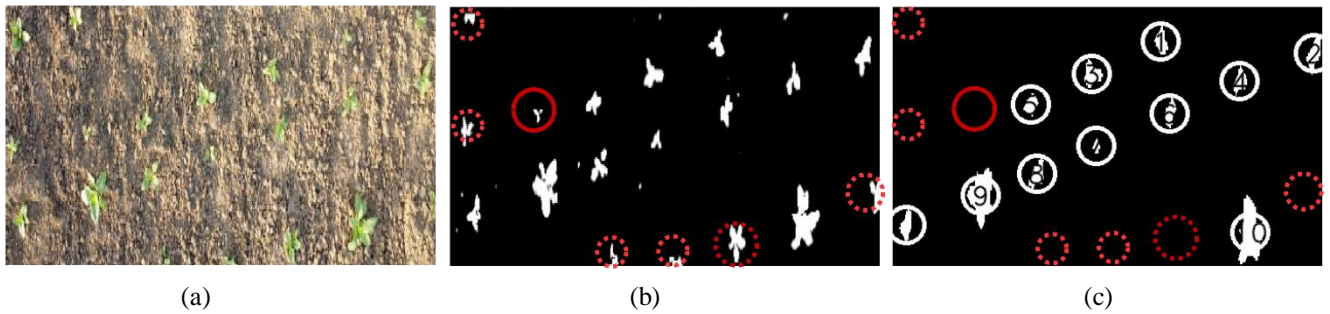


Fig. 7. Qualitative analysis of the proposed framework with semantic segmentation (a) Image after overlap region removal (Manually cropped the masked overlap region for better visualization); (b) U-Net prediction (Manual cropped overlap region from predicted image) (a); (c) count on the predicted image(b). Solid red circles indicate tobacco plants that are not counted as removed by morphological operations due to smaller sizes. Dotted red circles show the tobacco plants that are ignored as boundary plants in every next image to avoid double counting.

Input and predicted images after manually cropped overlap region for better visualization is shown in **Fig. 7a** and **Fig. 7b** whereas **Fig. 7c** is the result of proposed framework with semantic segmentation that includes cropping the identified overlap region followed by morphological operations and counting through connected pixel area.

The data presented in **Table I** clearly indicates that the estimated counts are consistently lower than the ground truth counts. This underestimation is due to two primary factors depicted in **Fig. 7**. Firstly, plants that are not included in the count due to their removal from the predicted images through morphological operations. Secondly, certain boundary plants are excluded from the count in both consecutive frames.

3.2 Object Detection

The model is trained on 83 images and validated on 20 images with a resulting mean average precision of 0.988. The tobacco detection with this trained object detection (YOLOv7) model is shown in **Fig. 6d**. Counting is done through counting the number of bounding boxes generated after detection.

Table II
Crop Emergence Estimation using Object Detection

Name	No. of images	True Count	Estimated	Precision	Recall	F1 Score
DATASET-1	28	472	460	1	0.97	0.987
DATASET-2	29	478	500	0.956	1	0.977
DATASET-3	32	522	535	0.975	1	0.988
DATASET-4	40	536	498	1	0.929	0.963
DATASET-5	61	540	459	1	0.85	0.918
Average	–	–	–	0.986	0.9498	0.9667

The evaluation of this proposed algorithm is shown in **Table II**. Ground truth, accuracy, precision, and recall are calculated as mentioned in the previous section. The evaluation of the proposed solution yielded an average F1 score of 0.9667, a precision score of 0.9852, and a recall score of 0.9484. The variance in the estimated count can be attributed to two factors. Firstly, when plants are present at the corners of consecutive frames, they may be counted twice, leading to higher estimated value than ground truth. This occurs because the considerable size of partially present plants in both frames can cause them to be counted as separate entities. Secondly, dataset having smaller plants leads to lesser estimated count as compared to ground truth. This is due to not identifying the very small tobacco plants. **Fig. 8** provides visual evidence of the two factors mentioned earlier on the same input as in **Fig. 7a**. In **Fig. 8b**, the solid blue circles represent the partial corner plants. It is evident that some of these plants have a significant portion of their structure present in the subsequent frame, leading to the possibility of double counting. Additionally, **Fig. 8b** also highlights the presence of smaller plants that are not classified as tobacco.



(a)

(b)

Fig. 8. Qualitative analysis of the proposed framework with objection detection (a) Image after overlap region removal (manually cropped the masked overlap region for better visualization); (b) prediction on the image (a) with object detection model (manually cropped overlap region from predicted image for better visualization) (b). Solid blue circles indicate tobacco plants that are counted using proposed framework with object detection but not with the proposed framework with semantic segmentation. The solid red rectangle shows the tobacco plant that is not identified as tobacco.

3.3 Object Detection & Tracking

Tobacco plant estimation using object detection with object tracking is shown in **Fig. 6e**. Unique identifications (IDs) can be seen on the bounding boxes, and the count is estimated by the total number of IDs generated. **Table III** shows the evaluation of this technique on video clips.

Table III
Crop Emergence Estimation using Real Time Applicable Framework

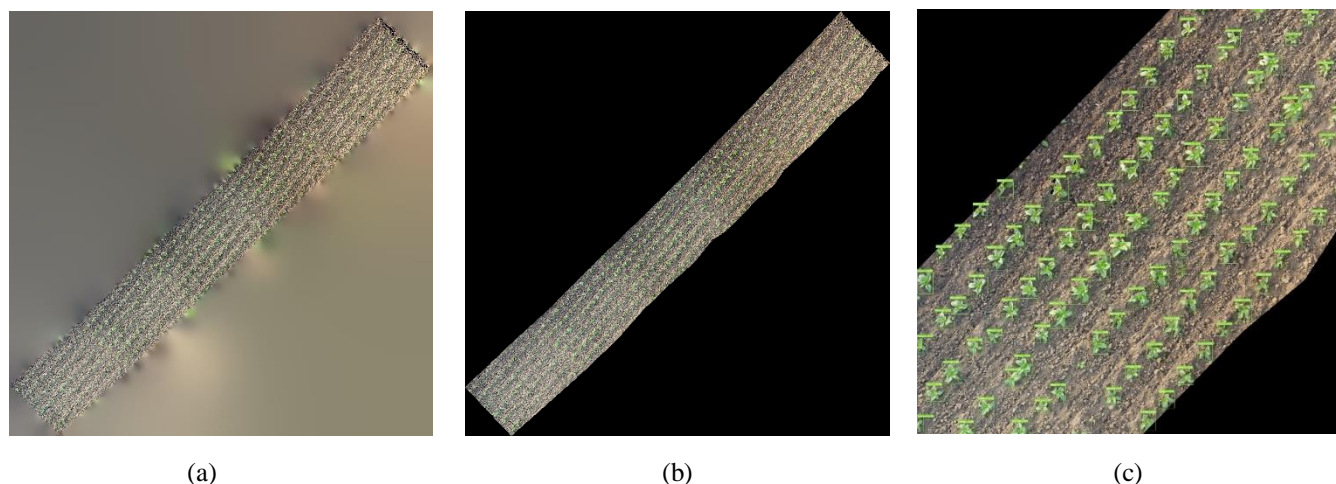
Name	True Count	Estimated	Precision	Recall	F1 score
DATASET-1	472	496	0.951	1	0.974
DATASET-2	478	518	0.922	1	0.959
DATASET-3	522	566	0.922	1	0.959
DATASET-4	536	560	0.957	1	0.978
DATASET-5	540	578	0.931	1	0.966
Average	–	–	0.9366	1	0.967

It can be observed from the results that it gives a count greater than the true value which is due to the switching IDs of the partial tobacco plants at the side rows of the video clips as occlusion and reappearance of the same plant is detected as a new entry. This is a promising methodology for real-time estimation with a recall score equal to 1 resulting in higher sensitivity for detection and good enough precision making it a promising approach for counting.

3.4 Orthomosaic-based Counting

We have also evaluated orthomosaic-based counting for highlighting time-efficiency of the proposed plant counting framework that involves: orthomosaic formation followed by patchifying (extraction of small images) this very high-resolution image to smaller images, plant detection on patches using previously discussed trained YOLOv7, and again reconstruction of large image from these plant-detected patches. Fig. 9. shows the orthomosaic of Dataset-1. Orthomosaic is formed using Agisoft Metashap (*Agisoft Metashape: Agisoft Metashape*, n.d.). Orthomosaic is preprocessed to remove blurred edges and resized for extraction of patches with the least overlap between patches. Reconstruction of large image from the detected patches is used to identify the plants that are counted twice due to the partial presence in more than one patch. The observed limitations include missing some frames during orthomosaic formation. These excluded frames can cause underestimation of manual counting plant in an orthomosaic.

The comparison of proposed techniques is presented in Fig. 10. revealing distinct patterns in the estimated plant counts. The semantic segmentation technique consistently underestimates the plant count, while the object detection and tracking plant counting framework consistently overestimates as observed from Fig. 10.



374

375

376

377

378

379

380

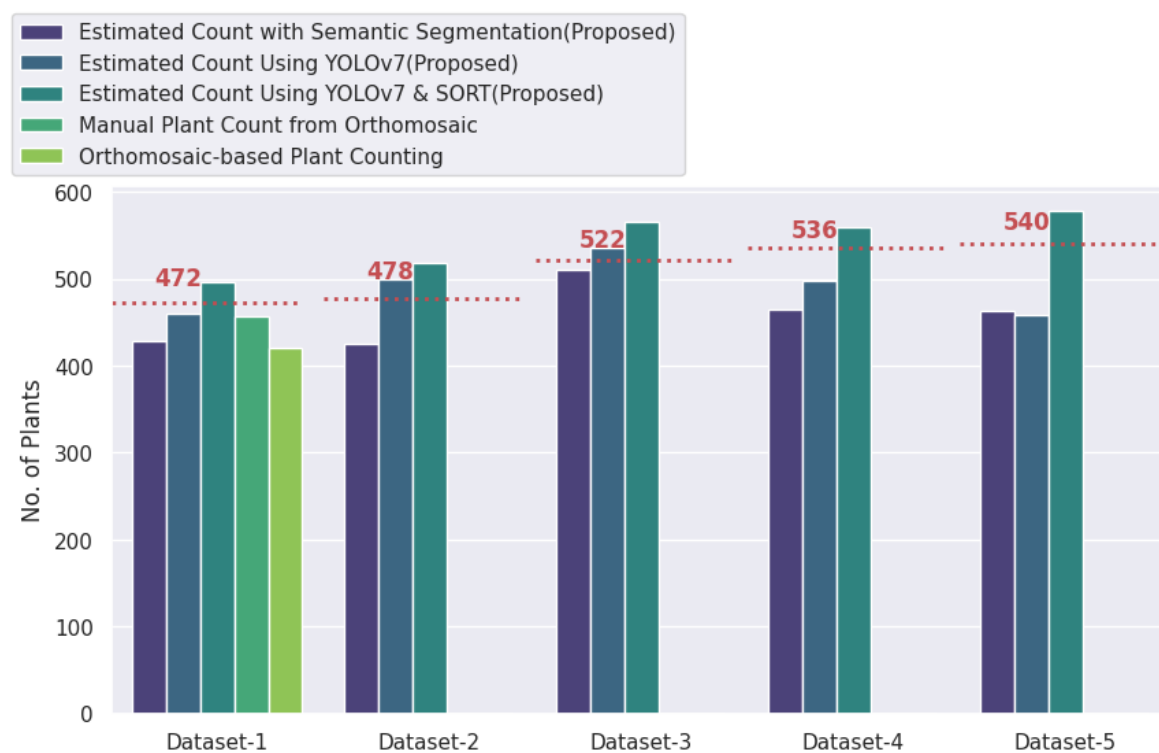
381

382

383

Fig. 9. Orthomosaic-based counting 7998×7998 (a) orthomosaic of DATASET-I; (b) cropped orthomosaic to remove blurred edges; (c) plant detection with YOLOV7 on a patch size of 1920×1920.

These observations suggest that both techniques are independent of the dataset's nature as all the datasets have varying tobacco plant sizes, soil and sunlight condition. In contrast, the object detection-based technique exhibits variance in the count, indicating its dependency on the specific characteristics of the dataset. Manual by hand counting plants in an orthomosaic results in a lesser count of 457 as compared to the observed ground truth of 472 from a video clip. This underestimation indicates one or more frames missed during time intensive orthomosaic formation and also due to blurred corners. The proposed object detection-based technique evaluation on the patches extracted from orthomosaic shows underestimation of with the count of 421 which is lesser compared to the proposed framework.



384

385

Fig. 10. Comparison of plant counting techniques. Dotted red lines show the ground truth for each dataset

Speed analysis of all the proposed techniques and orthomosaic-based technique is shown in Fig. 11. The real-time plant counting framework is evaluated on videos resulting in a higher average time. Results indicates that the proposed pipelines as shown in Fig. 1 require very less time and computational power as compared to the state-of-the-art techniques that involve orthomosaic formation.

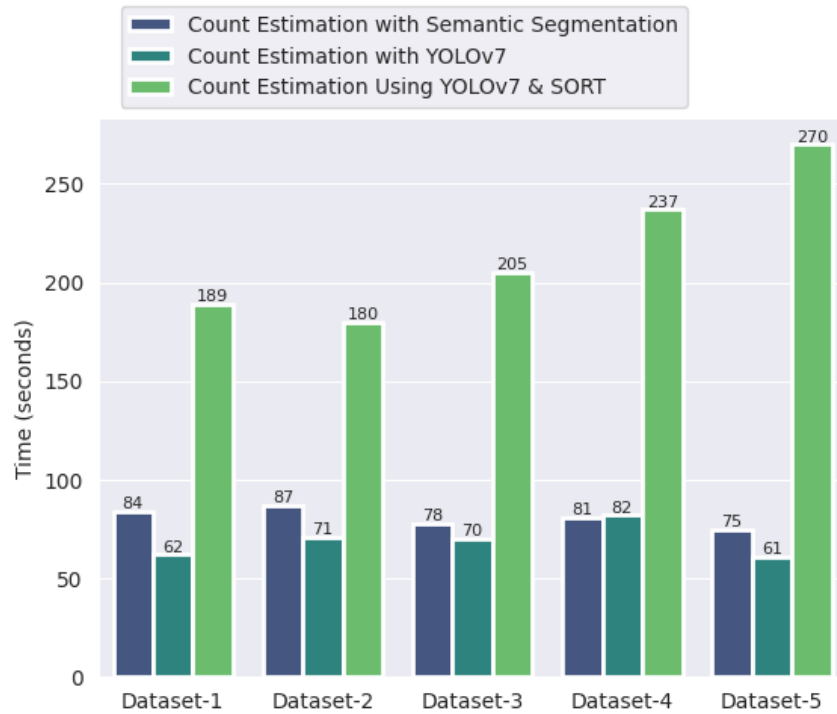


Fig. 11. Speed analysis of the proposed techniques.

The orthomosaic formation for DATASET-1 as shown in Fig. 9a. takes 6639s(1 hr and 50 min) whereas the proposed overlap detection algorithm on the same dataset with the same hardware specifications of 7.92 GB RAM, CPU Intel(R) Core(TM) i5-7200U, CPU 2.50GHz and AMD Radeon(TM) R5 M430 (Hainan) GPU takes 184s.

The proposed framework is implemented in Python programming. All the computations in this research are conducted using Google Colab with Intel Xeon CPU 2.2GHz, 12 GB RAM and Tesla K80 12 GB GPU.

4. CONCLUSION

In this study, we proposed a deep learning based framework for crop emergence estimation and conducted an evaluation of various potential solutions for the core modules. Specifically, we considered the utilization of U-Net as a plant detection module, as well as explored the effectiveness of using YOLOv7 as an alternative plant detection module. Furthermore, we have analyzed the combination of YOLOv7 with SORT to create a real-time framework for plant counting. The primary objective is to find a more efficient and accurate approach for tobacco plant counting, considering both time efficiency and computational requirements. Proposed framework with overlap detection module shows comparable accuracies for counting with the orthomosaic-based computationally expensive approaches. Overlap detection followed by U-Net plant detection module results in an average F1 score of 0.947. Overlap detection followed by YOLOv7 plant detection module results in 0.9667 average F1 score which is more than that of segmentation. A real-time system for plant counting through YOLOv7 and SORT was evaluated on recorded data but applicable to real-time resulting in 0.9672 average F1 score. However, critical evaluation of proposed framework highlights certain limitations including high variance and higher sensitivity of proposed

framework with YOLOv7 detection module that represents its dependency on the nature of dataset and counting some larger weeds as plant. Furthermore, U-Net based plant detection module results in underestimation of plant count due to pixel-level classification, and also ignorance of boundary objects in every next frame lead missing of some smaller plants at the boundary. Moreover, the detection and tracking approach shows ID switching for the partial plants at the corners of the video. Additionally, the collected aerial data have almost no weed infestation and the proposed algorithm might not show the same results for higher weed infestation. We evaluated the proposed framework on the datasets with only vertical overlap between frames, which corresponds to the one-directional motion of the UAV. In future, this proposed framework will be combined with the weed classification technique proposed by (Moazzam et al., 2022) and will be evaluated on other plants.

CONFLICT OF INTEREST

The authors have no conflicts of interest to disclose.

DATA AVAILABILITY STATEMENT

The data supporting this study's findings and analyzed during the current study will be made available from the corresponding author upon reasonable request.

ACKNOWLEDGEMENTS

We express our sincere appreciation and gratitude to the National Centre of Robotics and Automation (NCRA) for their invaluable contribution in providing the tobacco field data for our research. Their generous support has played a pivotal role in enabling us to carry out our study and achieve meaningful outcomes.

REFERENCES

- Agisoft Metashape: Agisoft Metashape.* (n.d.). Retrieved June 29, 2023, from <https://www.agisoft.com/>
- Alcantarilla, P. F., Bartoli, A., & Davison, A. J. (2012). KAZE features. *European Conference on Computer Vision*, 214–227.
- Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (n.d.). *SIMPLE ONLINE AND REALTIME TRACKING.* <https://github.com/abewley/sort>
- Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *ArXiv Preprint ArXiv:2004.10934.*
- Charlton, A. (2004). Medicinal uses of tobacco in history. *Journal of the Royal Society of Medicine*, 97(6), 292–296.
- Egi, Y., Hajyzadeh, M., & Eyceyurt, E. (2022). Drone-Computer Communication Based Tomato Generative Organ Counting Model Using YOLO V5 and Deep-Sort. *Agriculture (Switzerland)*, 12(9). <https://doi.org/10.3390/agriculture12091290>
- Fan, Z., Lu, J., Gong, M., Xie, H., & Goodman, E. D. (2018). Automatic Tobacco Plant Detection in UAV Images via Deep Neural Networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3), 876–887. <https://doi.org/10.1109/JSTARS.2018.2793849>
- Feng, A., Zhou, J., Vories, E., & Sudduth, K. A. (2020). Evaluation of cotton emergence using UAV-based imagery and deep learning. *Computers and Electronics in Agriculture*, 177(August), 105711. <https://doi.org/10.1016/j.compag.2020.105711>
- Fischler, M. A., & Bolles, R. C. (1981). RANSAC: Random Sample Paradigm for Model Consensus: A Application to Image Fitting with Analysis and Automated Cartography. *Graphics and Image Processing*, 24(6), 381–395.
- GitHub - heartexlabs/labelImg: LabelImg is now part of the Label Studio community. The popular image annotation tool created by Tzutalin is no longer actively being developed, but you can check out Label Studio, the open source data labeling tool for images, text, hypertext, audio, video and time-series data.* (n.d.). Retrieved January 7, 2023, from <https://github.com/heartexlabs/labelImg>

- 453 *GitHub - wkentaro/labelme: Image Polygonal Annotation with Python (polygon, rectangle, circle, line, point and*
454 *image-level flag annotation)*. (n.d.). Retrieved January 23, 2022, from <https://github.com/wkentaro/labelme>
- 455 Jeon, E., Kim, S., Park, S., Kwak, J., & Choi, I. (2021). Semantic segmentation of seagrass habitat from drone imagery
456 based on deep learning: A comparative study. *Ecological Informatics*, 66, 101430.
- 457 Jha, K., Doshi, A., Patel, P., & Shah, M. (2019). A comprehensive review on automation in agriculture using artificial
458 intelligence. In *Artificial Intelligence in Agriculture* (Vol. 2, pp. 1–12). KeAi Communications Co.
459 <https://doi.org/10.1016/j.aiia.2019.05.004>
- 460 Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. In *Computers and Electronics*
461 *in Agriculture* (Vol. 147, pp. 70–90). Elsevier B.V. <https://doi.org/10.1016/j.compag.2018.02.016>
- 462 Kitano, B. T., Mendes, C. C. T., Geus, A. R., Oliveira, H. C., & Souza, J. R. (2019). Corn Plant Counting Using Deep
463 Learning and UAV Images. *IEEE Geoscience and Remote Sensing Letters*, 1–5.
464 <https://doi.org/10.1109/lgrs.2019.2930549>
- 465 Li, Z., Guo, R., Li, M., Chen, Y., & Li, G. (2020). A review of computer vision technologies for plant phenotyping.
466 *Computers and Electronics in Agriculture*, 176, 105672. <https://doi.org/10.1016/J.COMPAG.2020.105672>
- 467 Liu, T., Wu, W., Chen, W., Sun, C., Zhu, X., & Guo, W. (2016). Automated image-processing for counting seedlings
468 in a wheat field. *Precision Agriculture*, 17(4), 392–406. <https://doi.org/10.1007/s11119-015-9425-6>
- 469 Maja, M. M., & Ayano, S. F. (2021). The Impact of Population Growth on Natural Resources and Farmers' Capacity to
470 Adapt to Climate Change in Low-Income Countries. In *Earth Systems and Environment* (Vol. 5, Issue 2, pp.
471 271–283). Springer Science and Business Media Deutschland GmbH. [https://doi.org/10.1007/s41748-021-](https://doi.org/10.1007/s41748-021-00209-6)
472 00209-6
- 473 Mancini, A., Frontoni, E., & Zingaretti, P. (2019). Satellite and uav data for precision agriculture applications. *2019*
474 *International Conference on Unmanned Aircraft Systems (ICUAS)*, 491–497.
- 475 Miao, C., Pages, A., Xu, Z., Rodene, E., Yang, J., & Schnable, J. C. (2020). Semantic Segmentation of Sorghum Using
476 Hyperspectral Data Identifies Genetic Associations. *Plant Phenomics*, 2020, 1–11.
477 <https://doi.org/10.34133/2020/4216373>
- 478 Moazzam, S. I., Khan, U. S., Qureshi, W. S., Nawaz, T., & Kunwar, F. (2022). Towards automated weed detection
479 through two-stage semantic segmentation of tobacco and weed pixels in aerial Imagery. *Smart Agricultural*
480 *Technology*, 100142.
- 481 Nee, C., Conway, L. S., Zhou, J., Kitchen, N. R., & Sudduth, K. A. (2021). Early corn stand count of different
482 cropping systems using UAV-imagery and deep learning. *Computers and Electronics in Agriculture*, 186(May),
483 106214. <https://doi.org/10.1016/j.compag.2021.106214>
- 484 Oh, S., Chang, A., Ashapure, A., Jung, J., Dube, N., Maeda, M., Gonzalez, D., & Landivar, J. (2020). Plant counting of
485 cotton from UAS imagery using deep learning-based object detection framework. *Remote Sensing*, 12(18), 2981.
- 486 Osco, L. P., dos Santos de Arruda, M., Gonçalves, D. N., Dias, A., Batistoti, J., de Souza, M., Gomes, F. D. G., Ramos,
487 A. P. M., de Castro Jorge, L. A., Liesenberg, V., Li, J., Ma, L., Marcato, J., & Gonçalves, W. N. (2021). A CNN
488 approach to simultaneously count plants and detect plantation-rows from UAV imagery. *ISPRS Journal of*
489 *Photogrammetry and Remote Sensing*, 174, 1–17. <https://doi.org/10.1016/j.isprsjprs.2021.01.024>
- 490 Pang, Y., Shi, Y., Gao, S., Jiang, F., Veeranampalayam-Sivakumar, A. N., Thompson, L., Luck, J., & Liu, C. (2020).
491 Improved crop row detection with deep neural network for early-season maize stand count in UAV imagery.
492 *Computers and Electronics in Agriculture*, 178(August), 105766. <https://doi.org/10.1016/j.compag.2020.105766>
- 493 Parico, A. I. B., & Ahamed, T. (2021). Real time pear fruit detection and counting using yolov4 models and deep sort.
494 *Sensors*, 21(14). <https://doi.org/10.3390/s21144803>
- 495 *Potential for Export / PTB*. (n.d.). Retrieved December 30, 2021, from <https://ptb.gov.pk/potential-export>
- 496 Rahmawati, D., Alfita, R., Ulum, M., & Murdianto, D. (2021). Tobacco Farming Mapping To Determine The Number
497 Of Plants Using Contour Detection Method. *E3S Web of Conferences*, 328, 4007.
- 498 Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection.
499 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.
- 500 Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. *Proceedings of the IEEE Conference on*
501 *Computer Vision and Pattern Recognition*, 7263–7271.
- 502 Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.
503 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.
- 504 Saleem, M. H., Potgieter, J., & Arif, K. M. (2021). Automation in Agriculture by Machine and Deep Learning
505 Techniques: A Review of Recent Developments. In *Precision Agriculture* (Vol. 22, Issue 6, pp. 2053–2091).
506 Springer. <https://doi.org/10.1007/s11119-021-09806-x>
- 507 Shirzadifar, A., Maharlooie, M., Bajwa, S. G., Oduor, P. G., & Nowatzki, J. F. (2020). Mapping crop stand count and
508 planting uniformity using high resolution imagery in a maize crop. *Biosystems Engineering*, 200, 377–390.
509 <https://doi.org/10.1016/j.biosystemseng.2020.10.013>

- 510 Tan, C., Li, C., He, D., & Song, H. (2022). Towards real-time tracking and counting of seedlings with a one-stage
511 detector and optical flow. *Computers and Electronics in Agriculture*, 193.
512 <https://doi.org/10.1016/j.compag.2021.106683>
- 513 Tareen, S. A. K., & Saleem, Z. (2018). A comparative analysis of sift, surf, kaze, akaze, orb, and brisk. 2018
514 *International Conference on Computing, Mathematics and Engineering Technologies (ICOMET)*, 1–10.
- 515 Valente, J., Sari, B., Kooistra, L., Kramer, H., & Mücher, S. (2020). Automated crop plant counting from very high-
516 resolution aerial imagery. *Precision Agriculture*, 21(6), 1366–1384. <https://doi.org/10.1007/s11119-020-09725-3>
- 517 Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art
518 for real-time object detectors. *ArXiv Preprint ArXiv:2207.02696*.
- 519 Wang, Y., Zhou, Z., Huang, D., Zhang, T., & Zhang, W. (2022). Identifying and Counting Tobacco Plants in
520 Fragmented Terrains Based on Unmanned Aerial Vehicle Images in Beipanjiang, China. *Sustainability*
521 *(Switzerland)*, 14(13). <https://doi.org/10.3390/su14138151>
- 522 Yang, H., Chang, F., Huang, Y., Xu, M., Zhao, Y., Ma, L., & Su, H. (2022). Multi-object tracking using Deep SORT
523 and modified CenterNet in cotton seedling counting. *Computers and Electronics in Agriculture*, 202.
524 <https://doi.org/10.1016/j.compag.2022.107339>
- 525 Zhang, J., Xie, T., Yang, C., Song, H., Jiang, Z., Zhou, G., Zhang, D., Feng, H., & Xie, J. (2020). Segmenting purple
526 rapeseed leaves in the field from UAV RGB imagery using deep learning as an auxiliary means for nitrogen
527 stress detection. *Remote Sensing*, 12(9), 1403.
- 528