The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2/W8-2024
8th International ISPRS Workshop LowCost 3D - Sensors, Algorithms, Applications, 12–13 December 2024, Brescia, Italy

Trait-focused low-cost and stereo-based 3D plant modeling for phenotyping via deep neural detection

Muhammad Wattad*, Sagi Filin

Mapping and Geo-Information Engineering, Technion – Israel Institute of Technology, Haifa, Israel
sabrym@campus.technion.ac.il, filin@technion.ac.il

Keywords: Plant phenotyping, Deep learning, Detection, Stereo imaging.

Abstract

Phenotyping, the measurement of plants physical traits, plays a pivotal role in advancing sustainable agricultural practices. Therefore, developing efficient, low-cost, means to generate such measures is vital. Though image based 2D-driven methods are commonly applied for that purpose due to their processing simplicity, it is clear that only 3D information can offer the necessary plant details. Notwithstanding, the generation of such data is challenged by the requirement to acquire a large set of images or the use of active sensors, which exhibit sensitivity to illumination and require lengthy acquisition campaigns. Consequently, 3D plant phenotyping is presently limited to controlled laboratory conditions and is hardly applied in actual growth setups. To address this shortcoming, this paper argues that by focusing on relevant plant traits, modeling can be simplified and the need for detailed plant reconstruction can be relieved. Accordingly, only a minimal set of images, specifically a stereo pair, can suffice for the reconstruction, thereby providing a low-cost sensing solution. To facilitate the reconstruction, we adapt an anchor-free detection deep neural network and integrate low- and high-level features to accurately detect our plant traits of interest. As the paper demonstrates our adapted network facilitates a robust 3D reconstruction of the entities of interest. Performance analysis demonstrates how our detection is reliable and accurate compared to standard anchor-free frameworks, translating to accurate reconstruction, as we validate against 3D plant scans.

1. Introduction

By the year 2050, Earth's population is projected to reach 9.8 billion, resulting in a 70% increase in global food demand (Tomlinson, 2013; Ford et al., 2019). This surge necessitates a corresponding increase in crop production and food supply (Davis et al., 2016). To support crop yield enhancement and its resistance to the changing environmental conditions, measurement of plants' traits and morphology (aka phenotyping, Zhang et al., 2023a) become crucial. Notwithstanding, deriving physical plant traits is challenging due to their dynamic, complex, and deformable geometry, which features structural discontinuities and comprises a mix of surface-dominant elements and elongated linear structures with numerous branching and offshoots. In addition, due to the plants elastic form, the effect of wind, as in greenhouses or field conditions, continuously distorts their shape, and in turn affects their reconstruction (Paturkar et al., 2021; Forero et al., 2022; Medic et al., 2023).

To model the complex plant form, it is customary to restrict the acquisition to laboratory environments and to utilize active sensing technologies or image-based reconstruction models (Miao et al., 2021; Wattad et al., 2024). While the first option can yield accurate and detailed structural representation, the acquisition may be slow and exhibit sensitivity to reflective surfaces, outdoor illumination, and environmental interference (Qi et al., 2019; Schunck et al., 2021; Liu et al., 2023). The use of the second option requires the acquisition of large image sets to reconstruct the plants 3D forms. Therefore, they exhibit sensitivity to environmental variability and field conditions, affecting completeness in modeling and demanding controlled environments (Paulus, 2019; Elnashef et al., 2019; Wu et al., 2020; Hu et al., 2024). As 3D phenotyping in actual growth conditions is challenged by present solutions, alternative reconstruction models that relieve the need for long acquisition time are warranted. In that respect, this paper proposes a new approach that requires only a stereo pair to reconstruct important plant traits. As such it significantly simplifies the acquisition, and as we demonstrate, the modeling. Our focus is on modeling plant nodes, internode distance, and length, all are important indicators of plant growth and stress factors (Sibomana et al., 2013; Lati et al., 2019). By reversing the conventional paradigm of reconstruction followed by interpretation, we simplify the modeling, and by adapting a neural detection architecture, we can focus the reconstruction on the relevant growth-related traits. Our results demonstrate improved performance compared to standard frameworks all while using lesser means. Our contributions are: *i*) an improved neural architecture that accurately detects key plant traits, thereby enabling efficient 3D modeling; *ii*) a computationally efficient approach that minimizes the complexity involved in feature extraction; *iii*) improved performance compared to existing anchor-free detection frameworks; *iv*) significant reduction in image acquisition and processing time, ensuring a viable 3D modeling.

2. Related work

Image-based reconstruction of plants' 3D shape is commonly approached using structure from motion multi-view stereo (SfM-MVS), where dozens of images are commonly acquired to model the structure of a single plant (Rose et al., 2015; Paulus, 2019; Shi et al., 2019; Li et al., 2020; Gong et al., 2021; Luo et al., 2022; Wattad et al., 2024). Elnashef et al. (2019) studied the minimum necessary number for accurate plant reconstruction, and concluded that about 60 images were required for complete architectural representation, but 20 sufficed for a tolerable, partial one. Focusing on leaf angle distribution, Qi et al.

* Corresponding author

(2019) needed three sets of stereo pairs to reconst their shape. Le Louëdec and Cielniak (2021) utilized stereo and time-of-flight cameras, mounted on a robotic platform, to localize and estimate the size and shape of soft fruits. Gong et al. (2021) traced rice growth using a rotating platform comprising of a USB camera and an active light source for structured light based modeling. Clearly, outdoor illumination and acquisition time made this solution impractical in actual field conditions. Li et al. (2022) developed a spinning platform equipped with two to four cameras, depending on plant dimensions, to extract phenotyping parameters. Between 80 to 160 images were acquired within a two minutes span to generate sufficiently dense 3D point clouds. Saeed et al. (2023) introduced the PeanutNeRF, a neural radiance field (NeRF) aimed at achieving 3D reconstruction of peanut plants. An extensive 360^{o} video footage was required to reconstruct the plant. Hu et al. (2024) demonstrated how hundreds of 4k-quality images acquired all around a plant yielded successful NERF-based shape reconstruction, but being slow in training and failing with an insufficient set of samples, this timely approach is impractical to model the geometry of the complex plants structure.

Alternative approaches considered active sensing systems to alleviate the need to process image data. Wu et al. (2019) employed a laser triangulation scanner to extract the plant skeleton and utilized 3D learning and clustering algorithms to estimate its morphological parameters. Schunck et al. (2021) developed a robotic arm equipped with a laser triangulation scanner, yielding 2.6M 3D points per plant, to quantify traits and track development. Despite the detail and accuracy it provided, specialized laboratory conditions were needed for its application. Using less expensive equipment, Liu et al. (2023) proposed a dual Kinect v2 camera set for rapid 3D reconstruction. As the cameras were symmetrically positioned on opposite sides of the plant, filtering and merging the two sets to yield a single point cloud became a challenge. Zhang et al. (2024) utilized terrestrial laser scanners to extract leaf-related phenotypical parameters. Accordingly, the proposed platform required a controlled lab environment and high processing resources for information extraction.

Due to the limitations of individual technologies, multi-sensor platforms have also evolved. Atefi et al. (2019) developed a robotic setup involving a broad range of sensors including active time-of-flight (TOF), RGB, spectral, and near-infrared (NIR) cameras, for modeling and extracting phenotypical parameters. Pérez-Ruiz et al. (2020) developed a field-based high-throughput phenotyping system that measures wheat canopy height and was built on a self-propelled robotic platform that navigates agricultural fields. The platform integrated LiDAR, for detailed 3D structural data capture, spectral sensors, for reference purposes, and an odometry sensors for accurate tracking and positioning. Interpretation of the data streams from this complex setup still required human interaction.

The review shows that current 3D reconstruction methods are resource-intensive, yet require long data acquisition campaigns and exhibit sensitivity to outdoor environmental conditions. Hence they are limited to laboratory environments and are impractical for use in field conditions. To address these challenges, the paper proposes a new reconstruction setup that in contrast to prevalent approaches requires only an instantaneously acquired stereo-pair to model essential plant-related traits in 3D. It focuses on the plant's length and internode distances, key phenotypical parameters to assess growth (Sibomana et al., 2013; Lati et al., 2019). As such our focus is on their extraction, but instead of relying on a reconstructed whole plant 3D model for that purpose, we target them directly in the image data. When successfully detected, even stereo pair would suffice for their reconstruction, thereby offering a low-cost readily available sensing device suitable for field applications. To materialize this concept, the paramount challenge lies in accurately detecting plant nodes, which we formulate via anchor-free object detection neural framework. Our analysis (Sec. 4) shows as node are small objects, detection could be inaccurate and inconsistent using common anchor-based and -free approaches (Girshick, 2015; Ren et al., 2016; Law and Deng, 2018; Carion et al., 2020; Jocher et al., 2022; Reis et al., 2023). Hence, our focus is on improving detection frameworks to support this challenge.

3. Methods

Our framework is anchor-free network-based, with the deep high-resolution convolutional neural network (HRNet, Sun et al., 2019) as its base architecture. The HRNet, designed for human pose detection through joints extraction, tends to perform more accurately and efficiently with small objects, compared to standard anchor-free counterparts (Law and Deng, 2018; Reis et al., 2023). As its naive application for plant node extraction exhibits high levels of misses and low recall (Sec. 4), we aim to modify and extend it to capture these intricate details.

In its core, our feature extraction module follows the base HRNet design of multiple parallel paths (aka branches) that process different resolutions, with up-sampling and fusion layers to integrate information from different resolutions (Fig. 1). The network graph consists of four sequential sub-networks (aka stages), each feeding the next. Its output is in the form of a heatmap (a 2D Gaussian) representing the localized key point entities in the data (Fig. 1). Although the base network focuses on fine image details, it still does not produce an informative enough feature map that can support the sought plant nodes extraction. To solve this shortcoming, we choose to enhance the representation of global context in the network with the aim of learning salient features that reflect the nodes uniqueness compared to their surroundings. To do so, we incorporate dual attention units (Fu et al., 2019) at each resolution branch and at the second to the last stage output. The role of these units is to trace, reflect, and enhance distinct and descriptive features globally and locally. Our dual attention comprises two subunits, *i*) a channel attention subunit that encourages distinct and descriptive features along channels, and *ii*) a spatial attention subunit, whose aim is to spatially locate salient regions. The input to both is a (same) feature map, $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels, and H and W are the respective height and width of the spatial dimensions. $\boldsymbol{F}$ is reshaped into $\boldsymbol{F}^{C}_{\text{res}} \in \mathbb{R}^{C\times(H\times W)}$, as input to the channel attention subunit. This allows us to specifically attend to distinct and descriptive features across channels, emphasizing differences in feature types. The product of $\boldsymbol{F}^{C}_{\text{res}}$ by its transpose, passed through a softmax function, generates our channel attention map, $\boldsymbol{M}_c$, which after standard attention normalization, generates our channel attention-derived feature map:

$$\boldsymbol{F}_{\text{channel}} = \boldsymbol{M}_c^{T}\boldsymbol{F} \tag{1}$$

For the spatial attention subunit $\boldsymbol{F}$ is passed through three convolutional layers to generate three new feature maps, $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{C\times H\times W}$. $\boldsymbol{Q}$ and $\boldsymbol{K}$ are reshaped into $\boldsymbol{Q}_{\text{res}}, \boldsymbol{K}_{\text{res}} \in \mathbb{R}^{(H\times W)\times C}$ to allow the attention subunit to focus on salient spatial regions

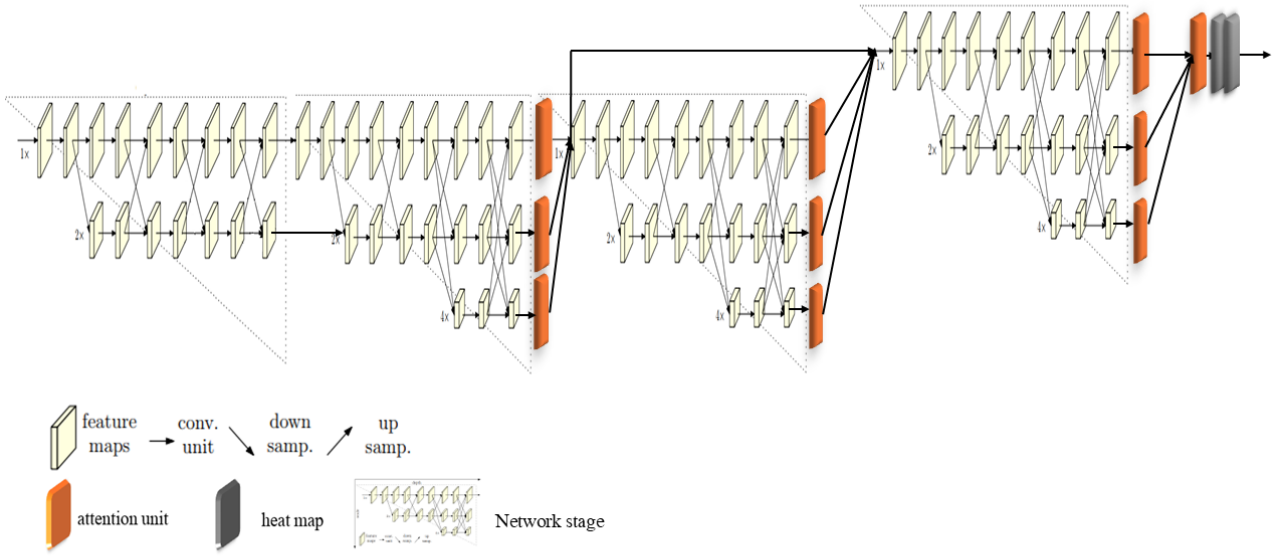

Figure 1. Our proposed architecture graph (adapted from Sun et al., 2019).

and locate important areas within the spatial dimensions. Generating the normalized spatial attention map, $\boldsymbol{M}_p^T$, the output is a weighted feature map for spatial positions,

$$\boldsymbol{F}_{\text{position}} = \boldsymbol{V}\boldsymbol{M}_p^T \tag{2}$$

where $\boldsymbol{V}$ acts as our values, and $\boldsymbol{F}_{\text{position}}$ retains the original $\mathbb{R}^{C\times H\times W}$ shape after applying the attention map. Finally, the enhanced feature map, $\boldsymbol{F}'$, combines both subunits output by:

$$\boldsymbol{F}' = \boldsymbol{F}_{\text{channel}} + \boldsymbol{F}_{\text{position}} \tag{3}$$

Notably, to concentrate all separately generated attended features, another dual attention unit is introduced at the last stage of the feature extraction (Fig. 1). Also, to integrate low- and high-level features into the saliency-related feature representation, and to enhance the network robustness to vanishing gradients, we introduce a skip connection from the output of the second stage to the input of the last, and concatenate the features therein.

Detection The heat maps, the product of the network, are generated by processing our output feature map through a detection head, essentially, a convolutional layer. This output was treated by the HRNet using a fixed number of channels (as human joints are fixed), each being directly tied to a specific joint. In such a manner, each channel consists of a single heatmap, located at the distinct joint placement. This has led the network to learn distinct human body joints, a property that is not shared by plant nodes. Additionally, and in contrast to human pose detection, here, the number of plant nodes may vary among species and between growth stages. To address these unique properties, we modify the output channels in the following manner, we consider the plant root as an individual entity and localize it using a designated channel, that is so because of its distinctiveness. Then, the plant nodes are all detected through a single channel that accommodates multiple heatmaps, related to the number of plant nodes. To detect the actual root position, the localization is carried out via non-maximum suppression. To detect the nodes, and as their number varies, we apply the mean-shift clustering algorithm (Comaniciu and Meer, 2002). Results show that our network handles the varying number of nodes effectively.

Loss function as the generated heatmaps are in the form of 2D Gaussians, the difference between the predicted output and the ground truth can be measured by the mean squared error loss, MSE $= \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, as in the base model. Ground truth heatmaps were generated as 2D Gaussians centered at the true node location with a standard deviation of 1 pixel. Thus, we enhance localization accuracy by encoding distance from the center, assigning higher values to points closer to the key location and progressively lower values farther away, thereby helping the model learn spatial importance and focus on precise point estimation.

3.1 3D localization and internode computation

The localization of our point of interest, allows us to utilize the epipolar constraint to both find correspondence and add another robustness layer to trace false alarm. Using the precomputed essential matrix, $\mathbf{E}_{3\times3}$, here with the baseline known to its actual dimension through calibration, correspondence should satisfy the point-to-line distance form:

$$\mathbf{x}^T\mathbf{E}\mathbf{x}'/\| \left(\mathbf{x}^T\mathbf{E}\right)_{1,2} \| \leq \varepsilon \tag{4}$$

and vice versa, where $\mathbf{x}$ and $\mathbf{x}'$ are the putative corresponding points in the two images, ε is the distance threshold to consider the two points as a match, and the subscripts in the denominator relate to the first two elements in the product $\mathbf{x}^T\mathbf{E}$. Defining the camera centers as $\mathbf{o}_i$, $i = [0, 1]$, and the image rays $\mathbf{v}_i = \mathbf{R}_i\mathbf{x}_i$, where $\mathbf{R}$ is the rotation matrix, the optimal triangulated object-

Figure 2. Performance analysis of our network against the YOLOv8 demonstrated as various growth stages, demonstrating our higher accuracy and reliability.

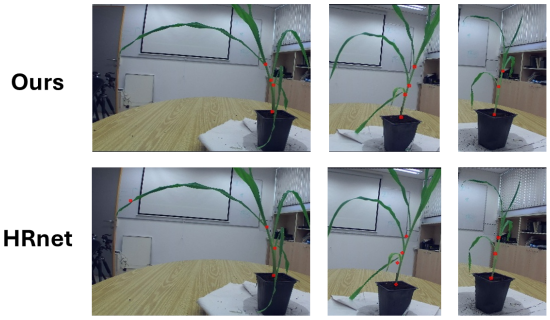

Figure 3. Performance analysis of our network against the HRNet, demonstrating higher accuracy and reduction in missed detections.

space point, $\mathbf{X}$, should minimize the objective function:

$$\arg\min_{\mathbf{X}} \sum_i \left(\mathbf{I} - \mathbf{v}_i \mathbf{v}_i^T\right) (\mathbf{X} - \mathbf{o}_i) \tag{5}$$

Notably, the plant height, the most distant point from the root, is extracted by tracing the green shades in the image. Obtaining the orientation of the plant by using the root point and the closest node to it, allows to trace it in reference to the plant growth direction. Detection of the furthest point, followed by epipolar correspondence and similar reconstruction allows us to compute it.

4. Results

We demonstrate our network performance on maize plants, an essential crop species known for its structural complexity and the absence of clear bifurcation points, a feature that sets it apart from species like tomato and cotton (Zhang et al., 2023b). This structural intricacy presents unique challenges for image analysis and makes maize a suitable subject for testing advanced phenotyping models. Our dataset comprises images of 47 maize plants acquired at different growth stages and exhibiting a wide range of morphological variations. The images were collected using a low-cost stereo ZED camera (Stereolabs, 2024), making the data acquisition process accessible and feasible for practical applications while still providing sufficient detail to assess our model's effectiveness under diverse growth conditions and plant forms. Moreover, to improve the generalizability and robustness of our model, the data underwent an extensive augmentation process, including rotation, clipping, and scaling.

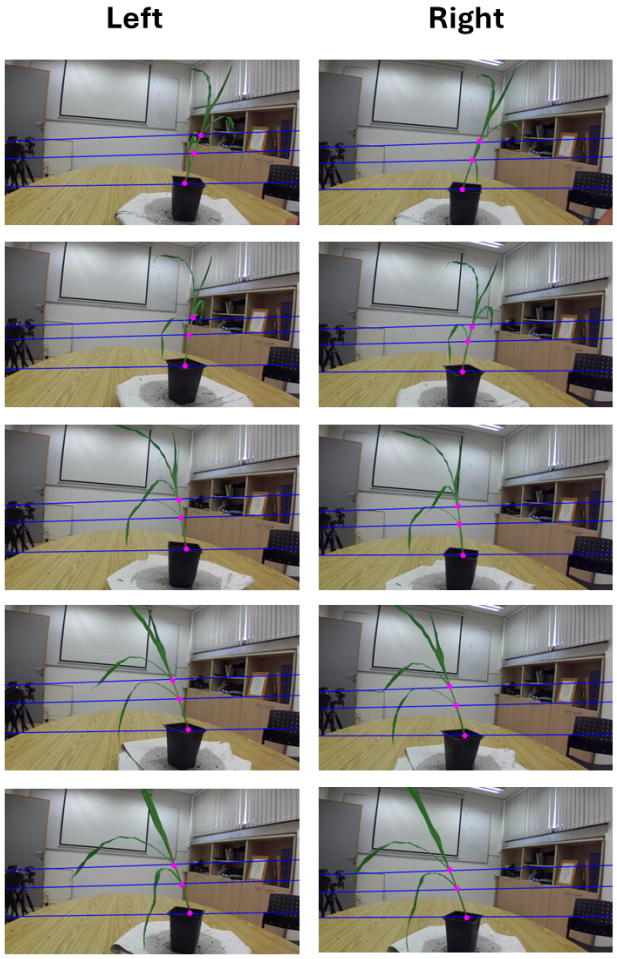

Figure 4. Stereo pair node detection, where in pink are the nodes and in blue epipolar lines.

We compared our model with the state-of-the-art YOLOv8 detection model (Reis et al., 2023), and the HRnet (Sun et al., 2019). Its performance is evaluated using the following metrics: precision, recall, and average pixel localization accuracy (APLE), which quantifies the mean distance between the predicted, P_i, and ground truth, G_i pixel locations, so that $E = \frac{1}{N}\sum_{i=1}^{N} \|P_i - G_i\|$, where N is the total number of evaluated nodes. For quantitative evaluation of the phenotyping-related measures, we compare the difference between the predicted object-space internode length, P_{3D}, and the corresponding ground truth values, G_{3D}, by $E_{3D} = \frac{1}{N}\sum_{i=1}^{N} \|P_{3D,i} - G_{3D,i}\|$.

Implementation details The network training was conducted on a single NVIDIA RTX A4000 GPU, utilizing PyTorch version 1.10.2 with CUDA 11.1. We trained the model for 500 iterations using the Adam optimizer, with a learning rate set to 0.001. To ensure a fair comparison of testing times, all learning-based methods were evaluated on the same NVIDIA RTX A4000 GPU. Additionally, all models were initialized using pre-trained weights from a general dataset and subsequently fine-tuned on our specific dataset to optimize performance.

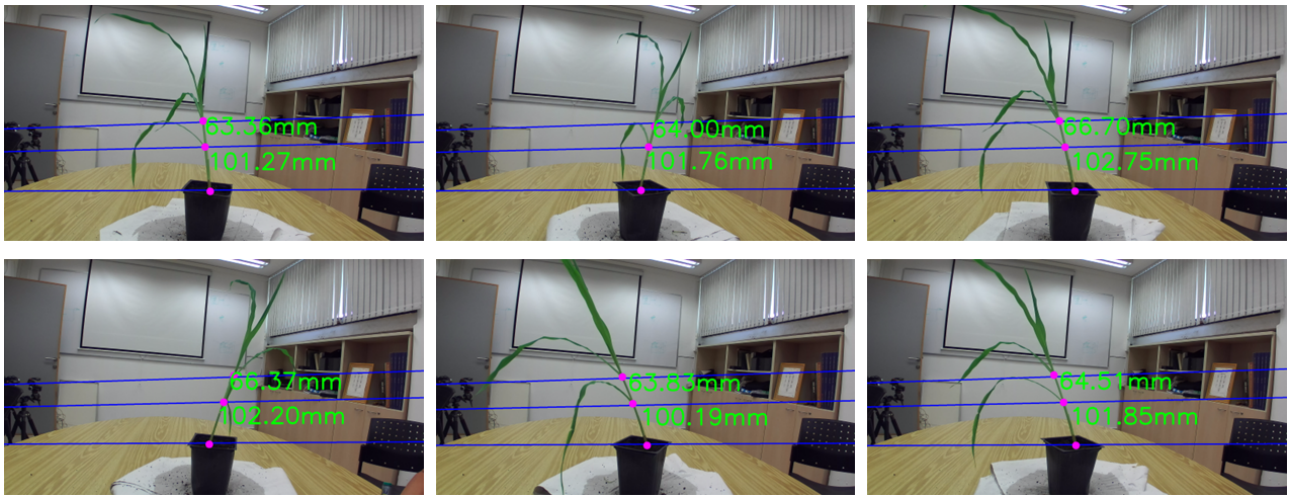

Figure 5. Computed internode lengths (green). Ground truth values are 99 mm (distance to the first node) and 63. 5mm.

Model	Prec.(%)	Recall(%)	Avg. Err. (Pix.)
Yolov8	85	71	21
Base NT	73	65	45
Ours	**91**	**89**	**6**

Table 1. Comparative analysis of precision, recall, and average error metrics for our network, YOLOv8, and HRNet.

4.1 Model analysis

Ablation study We examine the influence of our enhancements on the feature extractor. In Fig. (3), the left and middle samples represent the same plant from different viewpoints, observed 80 days after emergence. The samples contain three visible nodes. Unlike the base version, which exhibits notable inaccuracy in node localization, missed ones, and falsely detected non-existing others, ours accurately identifies all nodes correctly. This highlights the effectiveness of our feature extraction framework in achieving precise node detection. Table (1) also demonstrates how our model outperforms application of the commonly used YOLOv8 and the HRNet in terms of recall, precision, and detection error. A comparative analysis of our improved model and the HRNet equipped with our prediction head reveals the contribution of the attention units we introduced and shows how our enhanced architecture outperforms the baseline model in terms of detection accuracy and yielding a significantly lower rate of false alarms (Fig. 3).

Focusing on our framework performance against the Yolov8, Fig. (2), shows how the number of detected nodes changes with the images, e.g., two nodes in one and three in the other. Our in contrast consistently identifies the same, correct, number. This improved performance is attributed to our heatmap-based localization and feature extraction framework, which together enable precise node localization. Table (1) lists quantitative measures demonstrating the improved performance of our model compared to YOLOv8 in terms of recall, precision, and detection error. Our model also demonstrates higher accuracy in detecting and extracting the exact nodes at different growth stages while exhibiting a significantly lower rate of false alarms (Table 1). Its recall rate is 89% compared to 71% and 65% when using the Yolov8 and HRNet, respectively, with a pixel error of 6 pixels compared to 21 and 65 for the respective networks.

Quantitative evaluation Fig. (4), demonstrates the correctness of our detection and localization, manifested also through the accuracy by which they coincide with the epipolar line. These results apply to both nodes and the detected root, facilitating their actual 3D reconstruction. To test the accuracy of the metric object-space data, and as 3D models of the plant are available, we compared the internode distance by annotating key points in the 3D point cloud and computing the internode length as the geodesic distance as ground truth. As Fig. (5) demonstrates, our proposed method facilitates consistent metrics with errors not higher than 2 mm (equivalent to 2-3% error), illustrating the quality of our modeling results.

5. Conclusions

This paper presented a new learning-based framework for 3D plant reconstruction using a low-cost stereo camera and only a single image pair. By detecting the plant nodes in both images and utilizing epipolar geometry it offers simplicity and robustness. Realizing that the detection of these small, nearly indistinct entities in the overall image frame poses a challenge, a deep-learning framework has been developed to trace them. The paper demonstrated how the introduction of attention units and modification of the output layers facilitated both accuracy and robustness in detecting and localizing the varying number of nodes achieving high fidelity by learning salient features in an improved manner. Experiments confirm the efficacy of our model in both detection, reconstruction, and 3D plant traits measurement. Future research would focus on enhancing the model for a broader range of plant traits, and optimization that can lead to lesser computational load so that lower-end computational devices can be utilized to generate these measurements.

Acknowledgments

The authors would like to thank the Technion ecological garden staff, specifically Oren Azari – the garden agronomist, for the support and help extended with the preparation of plant material used for this research.

References

Atefi, A., Ge, Y., Pitla, S., Schnable, J., 2019. In vivo human-like robotic phenotyping of leaf traits in maize and sorghum in greenhouse. *Computers and Electronics in Agriculture*, 163, 104854.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. *European conference on computer vision*, Springer, 213–229.

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5), 603–619.

Davis, K. F., Gephart, J. A., Emery, K. A., Leach, A. M., Galloway, J. N., D'Odorico, P., 2016. Meeting future food demand with current agricultural resources. *Global Environmental Change*, 39, 125–132.

Elnashef, B., Filin, S., Lati, R. N., 2019. Tensor-based classification and segmentation of three-dimensional point clouds for organ-level plant phenotyping and growth analysis. *Computers and electronics in agriculture*, 156, 51–61.

Ford, N., Trott, P., Simms, C., 2019. Food portions and consumer vulnerability: qualitative insights from older consumers. *Qualitative Market Research: An International Journal*, 22(3), 435–455.

Forero, M. G., Murcia, H. F., Méndez, D., Betancourt-Lozano, J., 2022. LiDAR platform for acquisition of 3D plant phenotyping database. *Plants*, 11(17), 2199.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3146–3154.

Girshick, R., 2015. Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Gong, L., Du, X., Zhu, K., Lin, K., Lou, Q., Yuan, Z., Huang, G., Liu, C., 2021. Panicle-3D: efficient phenotyping tool for precise semantic segmentation of rice panicle point cloud. *Plant Phenomics*.

Hu, K., Ying, W., Pan, Y., Kang, H., Chen, C., 2024. High-fidelity 3D reconstruction of plants using Neural Radiance Fields. *Computers and Electronics in Agriculture*, 220, 108848.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., Fang, J., Yifu, Z., Wong, C., Montes, D. et al., 2022. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo*.

Lati, R. N., Filin, S., Elnashef, B., Eizenberg, H., 2019. 3-D image-driven morphological crop analysis: a novel method for detection of sunflower broomrape initial subsoil parasitism. *Sensors*, 19(7), 1569.

Law, H., Deng, J., 2018. Cornernet: Detecting objects as paired keypoints. *Proceedings of the European conference on computer vision (ECCV)*, 734–750.

Le Louëdec, J., Cielniak, G., 2021. 3D shape sensing and deep learning-based segmentation of strawberries. *Computers and Electronics in Agriculture*, 190, 106374.

Li, D., Shi, G., Kong, W., Wang, S., Chen, Y., 2020. A leaf segmentation and phenotypic feature extraction framework for multiview stereo plant point clouds. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 2321–2336.

Li, Y., Wen, W., Miao, T., Wu, S., Yu, Z., Wang, X., Guo, X., Zhao, C., 2022. Automatic organ-level point cloud segmentation of maize shoots by integrating high-throughput data acquisition and deep learning. *Computers and Electronics in Agriculture*, 193, 106702.

Liu, Y., Yuan, H., Zhao, X., Fan, C., Cheng, M., 2023. Fast reconstruction method of three-dimension model based on dual RGB-D cameras for peanut plant. *Plant Methods*, 19(1), 17.

Luo, L., Jiang, X., Yang, Y., Samy, E. R. A., Lefsrud, M., Hoyos-Villegas, V., Sun, S., 2022. Eff-3DPSeg: 3D organ-level plant shoot segmentation using annotation-efficient point clouds. *arXiv preprint arXiv:2212.10263*.

Medic, T., Bömer, J., Paulus, S., 2023. Challenges and recommendations for 3D plant phenotyping in agriculture using terrestrial lasers scanners. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 1007–1014.

Miao, T., Wen, W., Li, Y., Wu, S., Zhu, C., Guo, X., 2021. Label3DMaize: toolkit for 3D point cloud data annotation of maize shoots. *GigaScience*, 10(5), giab031.

Paturkar, A., Sen Gupta, G., Bailey, D., 2021. Making use of 3d models for plant physiognomic analysis: A review. remote sens. 13, 1–28.

Paulus, S., 2019. Measuring crops in 3D: using geometry for plant phenotyping. *Plant methods*, 15(1), 103.

Pérez-Ruiz, M., Prior, A., Martínez-Guanter, J., Apolo-Apolo, O. E., Andrade-Sanchez, P., Egea, G., 2020. Development and evaluation of a self-propelled electric platform for high-throughput field phenotyping in wheat breeding trials. *Computers and Electronics in Agriculture*, 169, 105237.

Qi, J., Xie, D., Li, L., Zhang, W., Mu, X., Yan, G., 2019. Estimating leaf angle distribution from smartphone photographs. *IEEE Geoscience and Remote Sensing Letters*, 16(8), 1190–1194.

Reis, D., Kupec, J., Hong, J., Daoudi, A., 2023. Real-time flying object detection with YOLOv8. *arXiv preprint arXiv:2305.09972*.

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6), 1137–1149.

Rose, J. C., Paulus, S., Kuhlmann, H., 2015. Accuracy analysis of a multi-view stereo approach for phenotyping of tomato plants at the organ level. *Sensors*, 15(5), 9651–9665.

Saeed, F., Sun, J., Ozias-Akins, P., Chu, Y. J., Li, C. C., 2023. Peanutnerf: 3d radiance field for peanuts. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6254–6263.

Schunck, D., Magistri, F., Rosu, R. A., Cornelißen, A., Chebrolu, N., Paulus, S., Léon, J., Behnke, S., Stachniss, C., Kuhlmann, H. et al., 2021. Pheno4D: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. *Plos one*, 16(8), e0256340.

Shi, W., van de Zedde, R., Jiang, H., Kootstra, G., 2019. Plant-part segmentation using deep learning and multi-view vision. *Biosystems Engineering*, 187, 81–95.

Sibomana, I., Aguyoh, J., Opiyo, A., 2013. Water stress affects growth and yield of container grown tomato (Lycopersicon esculentum Mill) plants. *Gjbb*, 2(4), 461–466.

Stereolabs, 2024. Zed stereo camera. `https://www.stereolabs.com`. Accessed: November 2, 2024.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5693–5703.

Tomlinson, I., 2013. Doubling food production to feed the 9 billion: a critical perspective on a key discourse of food security in the UK. *Journal of rural studies*, 29, 81–90.

Wattad, M., Alchanatis, V., Edan, Y., Shriki, S., Sandovsky, T., Filin, S., 2024. *Advances in the use of robotics in crop phenotyping*. 533–566.

Wu, S., Wen, W., Wang, Y., Fan, J., Wang, C., Gou, W., Guo, X., 2020. MVS-Pheno: a portable and low-cost phenotyping platform for maize shoots using multiview stereo 3D reconstruction. *Plant Phenomics*.

Wu, S., Wen, W., Xiao, B., Guo, X., Du, J., Wang, C., Wang, Y., 2019. An accurate skeleton extraction approach from 3D point clouds of maize plants. *Frontiers in plant science*, 10, 248.

Zhang, H., Wang, L., Jin, X., Bian, L., Ge, Y., 2023a. High-throughput phenotyping of plant leaf morphological, physiological, and biochemical traits on multiple scales using optical sensing. *The Crop Journal*.

Zhang, Q., Chen, Z., Zhou, Z., Wang, L., Liao, Q., Yang, C., Yang, J., 2024. 3D terrestrial LiDAR for obtaining phenotypic information of cigar tobacco plants. *Computers and Electronics in Agriculture*, 226, 109424.

Zhang, T., Elnashef, B., Filin, S., 2023b. Spatio-temporal registration of plants non-rigid 3-D structure. *ISPRS Journal of Photogrammetry and Remote Sensing*, 205, 263–283.