Tomato Fruit Evolution and Maturity estimation from RGB Time-lapse Observation via Modern Structure from Motion and Instance Segmentation

Stefan Baar, Yosuke Kobayashi, and Shinya Watanabe
Muroran Institute of Technology
Hokkaido, Japan
Email: {sbaar@, ykobayashi@, sin@}muroran-it.ac.jp

Abstract—In various cultivation environments, monitoring fruit development, such as phenotyping, is crucial to optimise and enhance plant and environmental conditions. A statistical approach is introduced in combination with computer vision to estimate tomato fruit maturity features, such as size and shape, from a set of sequential RGB images. An novel end-to-end routine has been implemented to locate and measure the size and colour of tomato fruit from handheld image sequences. This is achieved through a combination of instance segmentation and modern Structure from Motion (SfM) methods. Our approach is time-efficient and easy to implement in both small gardens and large greenhouse environments. Measuring tomato fruit size by hand, for example with a calliper, is limited by the symmetry of the fruit shape. In addition to estimating the fruit diameter, we can measure deviations from the targeted spherical topology with millimetre precision and incorporate any morphological abnormalities.

Index Terms—fruit maturity estimation, deep learning, instance segmentation, photogrammetry

I. Introduction

Monitoring fruit growth and quality is crucial to optimise environmental growing conditions to maximise the fruit yield at harvest. Here, estimating the fruit maturity plays a key factor. Passively (without distracting the growth process), the fruit maturity can be estimated through tomato size and colour measurements. Precisely estimating the fruit maturity for each fruit enables the estimation for time of harvest as well as fruit yield [1]. Estimating fruit size and morphology by hand can be difficult and time-consuming, especially for asymmetrically shaped fruits. Optical methods, including optical ranging and scanning, are often unsuitable due to the high reflectance and the light absorption characteristics of the fruit surface. However, advances in deep-learning based image segmentation and 3D-reconstruction (SfM, nerfs, gaussian splatting) make it possible to perform precise fruit measurements from a sparse set of 2D-RGB images [4], [17]. This paper proposes a scalable imaging pipeline that measures the fruit diameter and estimates its roundness (variation of symmetry from a spherical geometry). The individual pipeline components (Materials and Methods) will be explained subsequently. Similar studies have been conducted for other fruits, such as apples and cherries, using a stationary and supervised approach [2] [3]. Daily measurements manual and visual were performed for a set of three tomato fruits from anthesis to harvest. The results for both measuring approaches are presented and compared in the Results section. We evaluate and elaborate on limitations as well as future research in section: Discussion and Conclusion. Supplementary information, such as code, videos and data are available on Github (https://github.com/StefanBaar/Tomato-Evolution-SCIS).

II. Materials and Methods

For our experiment, we captured short videos of the same plant over a period of forty days. The videos focused on three tomato fruits and had a minimum duration of sixty frames. Due to the growth of the plant and the associated change in position of each tomato fruit on each day, the videos were taken from slightly different positions every time.

The experiments presented in this study were performed in an uncontrolled environment (outside), in proximity of the Muroran Institute of Technology Hokkaido, Japan (longitude 140°59′ E, latitude 42°19′ N). Muroran is located on the island of Hokkaido, which lies in a temperate climate zone and is the northernmost island of Japan. The area is classified as plant hardiness zone five. Temperatures in winter can reach values below −20 °C and are usually not above 30 °C in summer[1].

This study utilises three-dimensional reconstruction of each tomato fruit to estimate fruit size, using a familiar background pattern to calibrate the metric. The process is summarised in Figure 1 and elaborated in the following paragraphs.

A. Instance Segmentation

The publicly available Yolov8-seg model is used to perform panoptic segmentation and locate each fruit within individual frames. The model is not trained on tomatoes specifically. However, it was trained on instances that

[1]Muroran climate according to the weather park (https://ja.weatherspark.com)

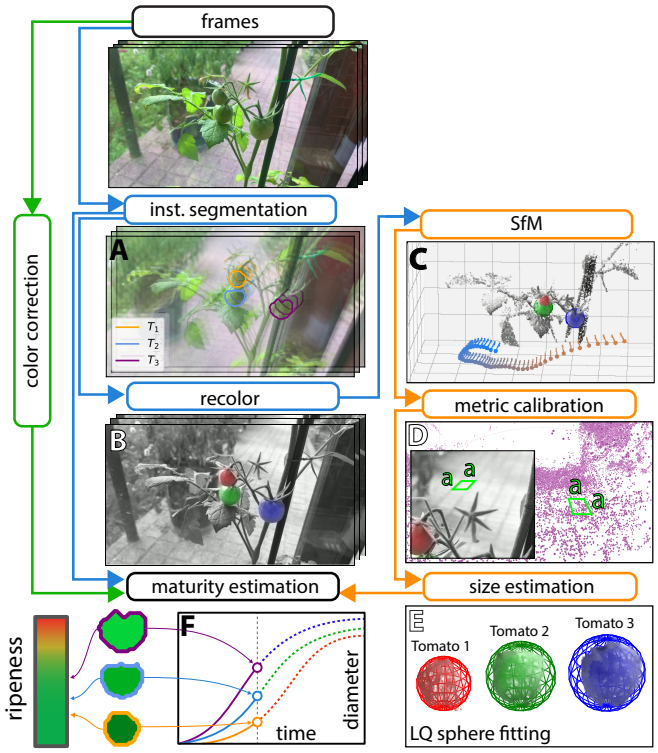

Fig. 1. Image processing pipeline: First, the tomatoes are are detected using instance segmentation (A). Following the images are dyed and tainted based on the individual tomato location (B). Next a 3D point cloud is computed using nerfacto, which is part of nerfstudio [4] (C). The point clouds are then calibrated using standard features (D). Following, the fruit morphology is analysed and the diameter evolution is evaluated using spherical regression (E). The individual fruit maturity is estimated via logistic regression (F).

include features that resemble round objects, as presented in Figure 2. The figure shows the activations of the last layer of the segmentation backend, inferred on the first frame of our dataset, as presented in Figure 1.

Backend Parameter Weighting is performed after apply the softmax function to the inference output (logits) of the model. The output probabilities $(p(g)_{m\times n})$ of the $N = 32$ individual groups n were combined using linear weights w(n) as presented in Eq. 1,

$$P_{tomato}^{u\times v} = \sum_{n}^{N} w(n)p(g)^{u\times v} \tag{1}$$

to produce a single set of probability maps $P_{tomato}^{u\times v}$ (retrieving tomato:0/background:1. The segmentations are then labeled (indexed) using connected components (2-connectivity). The individual labels are then sorted through out all connected frames using a Kalman filter-based sorting (tracking) approach. A direct comparison with the tracking methods provided by ultralytics will be provided elsewhere.

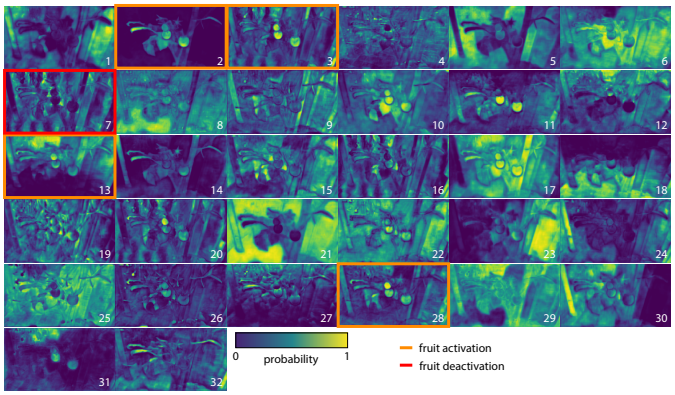

Fig. 2. YoloV8-seg Backend Activations: are presented after applying the softmax function to scale the outputs. Channels with significant activation and deactivation properties (separating the fruit from the background) are outlined in orange and red.

B. Image dying and fruit tainting (Colour Transformation)

Localising and identifying individual tomatoes in three dimensions is a major challenge. Objects located in the individual two-dimensional images $X \in \mathbb{I}^{(u\times v)}$ can be identified within a three-dimensional point cloud $Y \in \mathbb{R}^{(l\times 3)}$ by re-projecting the segmentation masks $\mathbf{M} \in \mathbb{I}^{(u\times v)}$ onto the point cloud. However, because of the SfM process filtering many image points (especially bundle adjustment), the map $f : X \mapsto Y$ describing the projection is a surjective map with $\forall y \in Y^3, \exists x \in X^2$.

$$\begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix} = \mathbf{RT} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \tag{2}$$

with $\mathbf{R}$ and $\mathbf{T}$ being the rotation and translation matrix with $\mathbf{R}, \mathbf{T} \in \mathbb{R}^{(3\times 3)}$. We assume, that precise projection of the 2D segmentations into the 3D point cloud can be challenging for complex morphologies as presented in our study by obstructions, e.g. leafs and other fruit [5] [6] [7]. Especially tomato leafs can take on complex shapes with high fractal dimensions f_D approaching $f_D > 2.5$. Therefore, instead of computing the projection of $\mathbf{M}' \in \mathbb{R}^{l\times 3}$ (l number of points) to $\mathbf{M} \in \mathbb{I}^{u\times v}$, the images are then re-coloured to make it possible to locate the individual fruits within the three-dimensional point cloud without the need to re-project segmentation masks. Here, we first transform the RGB into an greys-cale image and then assign individual colours to each tomato fruit by manipulating. Colours are applied by converting the RGB image into HSV image, where the first (H) , second (S) and third (V) channel represent colour, saturation and brightness, respectively. This means for 8bit images, a sparse set of 255 individual objects can be assigned by manipulating the first channel. This method holds the advantage that brightness and saturation are

preserved and the morphology of any of the tainted objects is unchanged (as presented in Figure 1 **B**). Note, that textures are preserved.

C. 3D Point clouds via Structure from Motion and Nerf

Point cloud are generated from the tainted images via costume and automated Nerfstudio [4] pipeline for each time-lapse observation. Image preprocessing, feature matching and initial SfM was performed using hloc [8]. The scenes are optimised using Neural radiance fields (nerfacto) [9]. The model is than exported as an oversampled point cloud containing lateral and colour information, along with its corresponding frame-dependent camera poses (intrinsic and extrinsic parameters).

D. Metric Scene Calibration

The produced point clouds are in the coordinate system inferred by the individual images. This means the unit/basis vectors $\mathbf{B} = (\vec{b}_1, \vec{b}_2, \vec{b}_3)$ are in relative coordinates resembling the pixel coordinates (u, v) of the individual images. This means the physical dimensions (in metres - m) cannot be inferred from the point clouds directly. Further, the coordinate system might be non-orthogonal in the form that the unit vectors are not equal in length ($||\vec{b}_1|| \neq ||\vec{b}_2|| \neq ||\vec{b}_3||$). This can cause embedded geometries to appear sheared. However, the coordinate system and its basis vectors can be calibrated (if sufficiently flat) by applying an appropriet coordinate transformation $\mathbf{F} : Y_{ref} \mapsto Y_{real}$ that maps the uncalibrated $Y_{ref} \in \mathbb{R}^{(l\times 3)}$ to the real metric coordinates $Y_{real} \in \mathbb{R}^{(l\times 3)}$. Here, the tensor $\mathbf{F}$ can be computed via solid reference object with well known morphology (e.g. spheres) or a flat repetitive pattern that is not orthogonal to any of the basis vectors. The This process is also known as Ground Control Points registration [10] [11].

E. Fruit identification and Sphere Fitting

The individual tomatoes are isolated (through staining as elaborated above) and approximated to estimate their diameter and morphological variance. The tomato species in this study is roughly spherical and does not exhibit strong elongations. In this study, we use Least Square Spherical regression [12], solving for the radius r and its projected centre of the sphere $Y^0 = [y_1^0, y_2^0, y_2^0]$ where the consolidated sphere (Eq. 3) is provided as:

$$||\vec{Y}||^2 = \mathbf{Y}'\vec{S}, \quad with \tag{3}$$

$$||\vec{Y}||^2 = \begin{bmatrix} y_{1,i}^2 + y_{2,i}^2 + y_{3,i}^2 \\ y_{1,i+1}^2 + y_{2,i+1}^2 + y_{3,i+1}^2 \\ \vdots \\ y_{1,J}^2 + y_{2,J}^2 + y_{3,J}^2 \end{bmatrix}, \tag{4}$$

$$\mathbf{Y}' = \begin{bmatrix} 2y_{1,i} & 2y_{2,i} & 2y_{3,i} & 1 \\ 2y_{1,i} & 2y_{2,i} & 2y_{3,i} & 1 \\ & \vdots & & \\ 2y_{1,J} & 2y_{2,J} & 2y_{3,J} & 1 \end{bmatrix}, \tag{5}$$

$$and \quad \vec{S} = \begin{bmatrix} y_1^0 \\ y_2^0 \\ y_3^0 \\ r^2 - (y_1^0)^2 - (y_2^0)^2 - (y_2^0)^2 \end{bmatrix} \tag{6}$$

F. Maturity Estimation

From the inferred diameter evolution produced by estimating the radius for each tomato of each observation, it is possible to fit the fruit diameter over time relationship. The fruit growth follows in general the characteristics of biological cell proliferation, which exhibits logistic growth over time x that can be described by the Richards function, as shown in Eq. 7. The function can be simplified by setting $c = 1$ and $b = z$.

$$f_R(x) = D_{tomato}(x) = a + \frac{A - a}{(c + be^{-d(x-x_0)})^{1/z}} \tag{7}$$

Further the fruit maturity $M_f(D_f)$ is the normalised degree of ripeness based on $D_h = D_f(0.95A)$ at the projected saturation of growth. Depending on the fruit diameter the relative fruit Maturity can be written as

$$M_f(D_h = 0.95A) = \frac{x}{\frac{1}{bd} ln(1.05^d - 1) + x_0}, \tag{8}$$

with $[0 \leq M_f \leq 1] \in \mathbb{R}$. Determining the regression variable b,d, and x_0, the fruit Maturity M can be computed with a single diameter measurement. Further, it is possible to estimate the time until full maturity for an initial set of measurements.

III. Results

The instance masked were produced with a mean Average Precision (mAP) of 0.7 considering an Intersection over Union (IOU) of 0.8 ,which is comparable to related studies. However, the training and evaluation of our segmentation approach will be discussed elsewhere. One sample frame of each observation is presented in Figure 3. Where day:0 is the first observation (random timing) taken after fruit anthesis. The time of fruit anthesis was determined through none-linear curve regression as can be derived from Equation 7

Evaluating the precision of a point cloud mapping real-world objects is cahllenging, since producing rational reference data is unfeasible in the context of this research. However, the validation error (image model re-projection rate) when training the nerf model (50/50 split of frames for training and validation) was estimated to be larger than 0.99. The tainted point cloud of three tomatoes and

Fig. 3. Fruit detection and tracking results:. Segmentation results for 27 sample observations. The same three tomatoes (indicated in orange, blue and purple) are tracked through all observations. The growing period for the selected species of tomatoes is about 40 days.

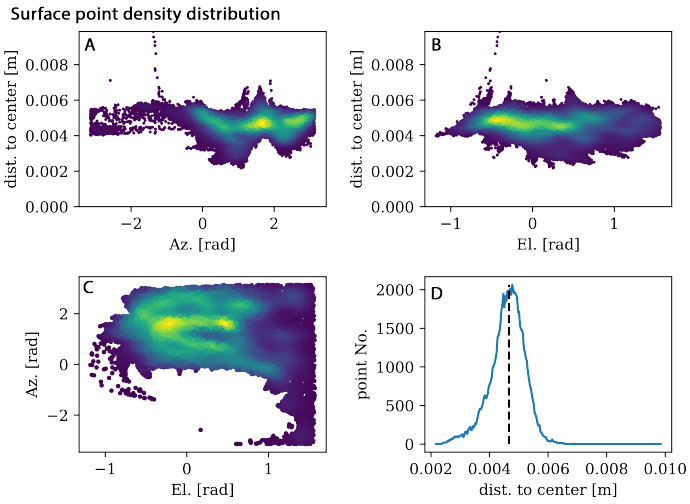

Fig. 5. Fruit morphology evaluation: The point density in spherical coordinates (Azimuth (A), Elevation (B) and distance to the centre of the sphere fitted to Tomato 1 (corresponding to Figure 1). C: Surface coverage and density. D: Radial point density histogram with strong peak at 4.5 mm.

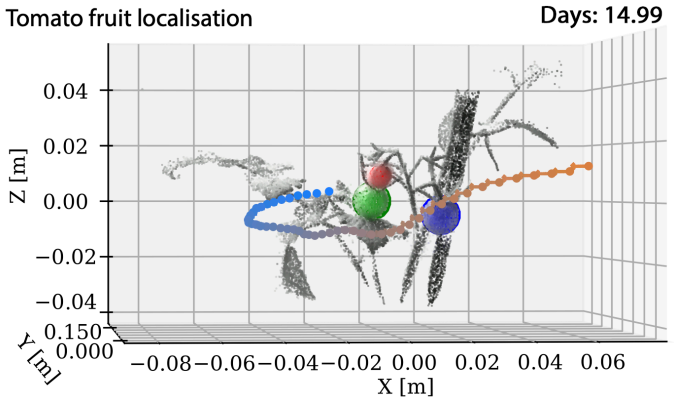

Fig. 4. 3D point cloud: of observation on day 14. Plant structures are in black and white, while tomatoes are highlighted in colour and spherical fits are indicated in blue (outline). The camera trajectory is indicated with arrows ranging from orange to blue (1/24 frames/s).

their environment is presented in Figure 4. The orange and blue arrows indicate the camera position and direction. The radial surface point distribution of Tomato 1 is displayed in 5. The distance of the individual data points associated with Tomato to the centre of the projected sphere is presented in Figure 5 A and B for the azimuth and elevation of sphere, respectively. The radial point distribution is almost normal distributed with a standard deviation of one millimetre. This means that the shape of the tomato 1 varies from that of sphere by approx. one millimetre and that the diameter of the fruit at the time of the measurement was estimated to be $d_1 = (9 \pm 1)mm$ as presented in Figure 5 D.

In Figure 6, we present the fruit diameter evolution over time for the three tomatoes indicated in Figure 3. The coloured dotes represent the diameter of each tomato measured through SfM. The grey dots represent the ground truth, where the individual tomato was measured with a caliber. Logistic fits were produced using the Richards curve for better comparability between the approaches. The logistic fits are indicated as solid lines in their respective colours.

IV. Discussion and Conclusion

In this study, we demonstrate that the tomato diameter can be sufficiently evaluated using modern methods for image segmentation and photogrammetry to produce robust results for varying experimental conditions. However, automated and precise fruit size measurements require a tight integration of the individual data-processing steps such as fruit detection, tracking and 3D reconstruction from a set of 2D images. While state of the art Structure from Motion (SfM) is robust, image segmentation and tracking remains the main bottleneck. Even though there are countless studies on tomato detection, they mostly utilise the standard segmentation or detection pipelines (e.g. any version of yolo), which are easy to implement and resource efficient, but limited in precision [13] [14] [15] [16]. Further, challenges arise when calibrating the world coordinates (pixel) from camera into real-world coordinates (metres). Reference objects or Ground Control Points (GCP) have to be established or identified for each observation. We will further discuss the influence of image quality as well as the choice of reference features for the metric calibration in future studies.

The estimated measurement uncertainties for both methods are $\Delta d_{i/m} = 1mm$. When fitting both measurements with a logistic functions, we confirm that the results (variation between regression curves) for both measurements only deviate by less than ten percent, even

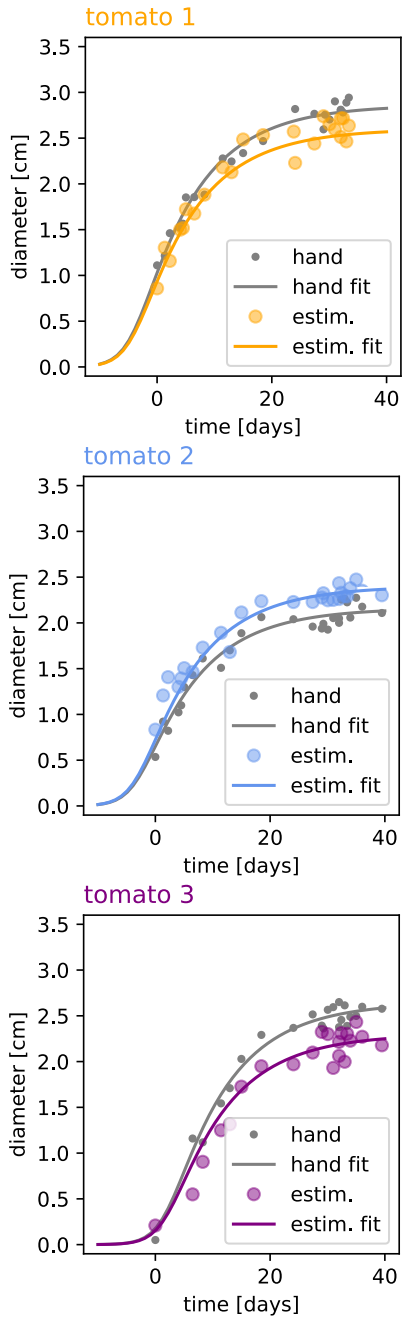

Fig. 6. Fruit diameter time evolution: The evolution of the fruit size over time for three fruits for three tomato fruits. The computed diameters are coloured, while the reference values are highlighted in grey. The measurement precision for both approaches are $\Delta d_{i/m} = 1\,\text{mm}$

though the variance from the regression curve is significant ($R^2_{hand} = 0.85 \pm 0.01, R^2_{estim} = 0.78 \pm 0.01$). The growth model and its properties have been elaborated in previous studies. This deviation is likely introduced by the irregular tomato shape that cannot be evaluated by manually measurements with a calliper for obvious reasons.

In our experiments, we have noticed radial distortions can distort the calibration results, Therefore, in future studies, we will evaluate additional SfM and photogrammetry approaches such as gaussian splatting [17].

Acknowledgments

This research was commissioned and supported by the NICT, JAPAN, and A-STEP of JST (Grant Number: JPMJTM20A1). This work was also supported in part by JSPS KAKENHI (Grant Number No. 20K11968 and No. 22H02463).

References

[1] S. Baar, Y. Kobayashi, T. Horie, K. Sato, and S. Watanabe, "Tomato fruit maturity estimation from rgb images," in 2022 IEEE 11th Global Conference on Consumer Electronics (GCCE). IEEE, 2022, pp. 615–616.

[2] J. Gené Mola, R. Sanz Cortiella, J. R. Rosell Polo, A. Escolà i Agustí, and E. Gregorio López, "Apple size estimation using photogrammetry-derived 3d point clouds," in 9th Annual Catalan Meeting on Computer Vision. September 19, 2022, Universitat Autònoma de Barcelona,(http://acmcv. cat/), 2022.

[3] J. Rong, Y. Yang, X. Zheng, S. Wang, T. Yuan, and P. Wang, "Three-dimensional plant pivotal organs photogrammetry on cherry tomatoes using an instance segmentation method and a spatial constraint search strategy," Available at SSRN 4482155, 2023.

[4] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," in ACM SIGGRAPH 2023 Conference Proceedings, ser. SIGGRAPH '23, 2023.

[5] M. Kellner, B. Stahl, and A. Reiterer, "Fused projection-based point cloud segmentation," Sensors, vol. 22, no. 3, p. 1139, 2022.

[6] J. Yang, C. Lee, P. Ahn, H. Lee, E. Yi, and J. Kim, "Pbp-net: point projection and back-projection network for 3d point cloud segmentation," in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 8469–8475.

[7] L. Kneip, D. Scaramuzza, and R. Siegwart, "A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation," in CVPR 2011. IEEE, 2011, pp. 2969–2976.

[8] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in CVPR, 2019.

[9] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," Communications of the ACM, vol. 65, no. 1, pp. 99–106, 2021.

[10] M. Pollefeys, "Self-calibration and metric 3d reconstruction from uncalibrated image sequences," Ph.D. dissertation, PhD thesis, ESAT-PSI, KU Leuven, 1999.

[11] C. G. Morton, J. L. Huntington, G. M. Pohll, R. G. Allen, K. C. McGwire, and S. D. Bassett, "Assessing calibration uncertainty and automation for estimating evapotranspiration from agricultural areas using metric," JAWRA Journal of the American Water Resources Association, vol. 49, no. 3, pp. 549–562, 2013.

[12] C. F. Jekel, Digital Image Correlation on Steel Ball, 2016. [Online]. Available: https://hdl.handle.net/10019.1/98627

[13] P. Wan, A. Toudeshki, H. Tan, and R. Ehsani, "A methodology for fresh tomato maturity detection using computer vision," Computers and electronics in agriculture, vol. 146, pp. 43–50, 2018.

[14] G. Liu, J. C. Nouaze, P. L. Touko Mbouembe, and J. H. Kim, "Yolo-tomato: A robust algorithm for tomato detection based on yolov3," Sensors, vol. 20, no. 7, p. 2145, 2020.

[15] M. O. Lawal, "Tomato detection based on modified yolov3 framework," Scientific Reports, vol. 11, no. 1, pp. 1–11, 2021.

[16] M. Zhu, "Yolov8 farm tomato detection based on attention mechanism and enhanced feature pyramid network," in International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2023), vol. 13105. SPIE, 2024, pp. 423–426.

[17] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," ACM Transactions on Graphics, vol. 42, no. 4, July 2023. [Online]. Available: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/