This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Wheat3DGS: In-field 3D Reconstruction, Instance Segmentation and Phenotyping of Wheat Heads with Gaussian Splatting

Daiwei Zhang[1*] Joaquin Gajardo[1*†] Tomislav Medic[1] Isinsu Katircioglu[2]
Mike Boss[1] Norbert Kirchgessner[1] Achim Walter[1] Lukas Roth[1]

[1]ETH Zürich [2]Swiss Data Science Center [*]Equal contribution [†]Corresponding author: jgajardo@ethz.ch

Webpage: https://zdwww.github.io/wheat3dgs/

Abstract

Automated extraction of plant morphological traits is crucial for supporting crop breeding and agricultural management through high-throughput field phenotyping (HTFP). Solutions based on multi-view RGB images are attractive due to their scalability and affordability, enabling volumetric measurements that 2D approaches cannot directly capture. While advanced methods like Neural Radiance Fields (NeRFs) have shown promise, their application has been limited to counting or extracting traits from only a few plants or organs. Furthermore, accurately measuring complex structures like individual wheat heads—essential for studying crop yields—remains particularly challenging due to occlusions and the dense arrangement of crop canopies in field conditions. The recent development of 3D Gaussian Splatting (3DGS) offers a promising alternative for HTFP due to its high-quality reconstructions and explicit point-based representation. In this paper, we present Wheat3DGS, a novel approach that leverages 3DGS and the Segment Anything Model (SAM) for precise 3D instance segmentation and morphological measurement of hundreds of wheat heads automatically, representing the first application of 3DGS to HTFP. We validate the accuracy of wheat head extraction against high-resolution laser scan data, obtaining per-instance mean absolute percentage errors of 15.1%, 18.3%, and 40.2% for length, width, and volume. We provide additional comparisons to NeRF-based approaches and traditional Muti-View Stereo (MVS), demonstrating superior results. Our approach enables rapid, non-destructive measurements of key yield-related traits at scale, with significant implications for accelerating crop breeding and improving our understanding of wheat development.

1. Introduction

Accurate and rapid measurement of plant traits is essential for advancing crop breeding programs and understanding plant development [2]. In particular, the precise characterization of wheat head morphology—including length, width, and volume—is critical for assessing yield potential and phenotypic variation of this staple crop [19]. Plant phenotyping traditionally relies on laborious manual measurements. Capturing and analyzing plant point clouds obtained with laser scans [26, 31, 55] and Multi-View Stereo (MVS) [9, 18, 24, 36, 57] provide automated alternatives, but are either too costly, slow or lack the fine-grained details required for accurate morphological measurements. Implicit neural representations, such as Neural Radiance Fields (NeRFs) [33, 34, 52], have emerged as a promising alternative for image-based plant phenotyping by overcoming limitations of traditional MVS approaches. By modeling space continuously, neural representations enable a more detailed reconstruction of intricate plant structures [3] and are better equipped to represent occluded regions and recover fine-grained details. Despite these advantages NeRFs remain computationally expensive, and require dense sampling of a neural network to obtain point

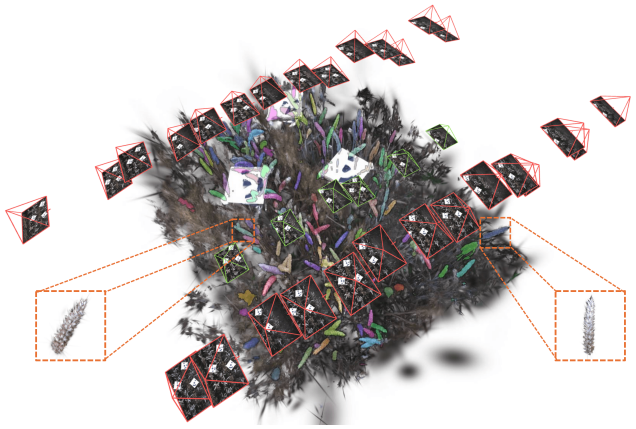


Figure 1. 3D Gaussian Splatting reconstruction of a wheat plot with segmented 3D wheat heads instances (in different colors). We use 30 views for reconstruction (red frustums) and 6 out-of-distribution views for evaluation (green frustums).

clouds [43], making editing and post-processing cumbersome.

While NeRFs have been applied in field conditions, their use has been limited to small-scale studies as proof-of-concept approaches or to measure simple traits on a handful of plants [42, 58]. Recent advancements in point-based 3D representations, particularly 3D Gaussian Splatting (3DGS) [21], offer a more efficient alternative of radiance fields for 3D reconstruction. 3DGS directly represents the scene with explicit Gaussian primitives, enabling fast rasterization and a more straightforward extraction of geometric features without additional post-processing. Although 3DGS has recently begun to be explored for plant phenotyping, its applications have so far been limited to controlled indoor environments at the single-plant level [37, 47]. Meanwhile, high-throughput field phenotyping (HTFP) is essential to support crop breeding programs and for yield prediction, but significant challenges arise in outdoor field conditions, such as heavy occlusions caused by other leaves or wheat heads in dense canopies. Even simple but critical tasks such as accurately detecting and counting wheat heads remain challenging [10], and common practice is still to do this manually.

To address these challenges and push the current capabilities, we propose Wheat3DGS, a novel pipeline that leverages 3DGS and the Segment Anything Model (SAM) [23] for 3D reconstruction of outdoor wheat canopies and 3D wheat head instance segmentation, enabling their individual extraction and measurement (Fig. 1). To address the challenge of segmenting individual wheat heads, we obtain per-view segmentation masks by prompting SAM with bounding boxes provided by an off-the-shelf wheat head detector—the winning model[1] of the 2021 Global Wheat Head Detection Challenge [10]. These masks are associated in 3D by projecting them onto the Gaussian splats, enabling the annotation and extraction of individual wheat heads in 3D space as groups of Gaussians. This hybrid approach combines the efficiency of explicit 3D representations with the semantic precision of advanced 2D vision models, facilitating accurate organ-level trait extraction.

We validate our method through comprehensive evaluation of canopy reconstruction quality, wheat head segmentation, and trait measurement quality. Our results demonstrate that Wheat3DGS outperforms NeRF-based in canopy reconstruction and 3D segmentation abilities, and exceeds MVS in wheat head trait measurement quality. Our main contributions can be summarized as follows:

- We show that 3D Gaussian Splatting can be effectively used for creating detailed 3D reconstructions of crop canopies from overhead RGB imagery and provide a quantitative and qualitative comparison to NeRF-based methods.
- We propose a method to perform 3D wheat head instance segmentation on a 3D reconstructed canopy by leveraging a pretrained wheat head detector and SAM, and associating semantic information in 3D. We also provide quantitative evaluation for extracted traits, and compare against high-resolution laser scan data, demonstrating superior accuracy and efficiency compared to MVS.
- We release a dataset comprising RGB images with calibrated camera poses, corresponding laser scans for seven wheat plots, and the view-consistent segmentation masks generated by our approach, providing a valuable resource for future research in image-based plant phenotyping.

By combining state-of-the-art 3D reconstruction methods with advanced segmentation techniques, Wheat3DGS addresses key challenges in automated plant phenotyping, improving our ability to measure wheat head morphology, and lays the foundation for future large-scale studies on crop development and yield prediction.

2. Related Work

3D reconstruction. Traditional 3D reconstruction methods such as Structure-from-Motion (SfM) [44, 49, 54] and MVS [15, 45, 56, 59, 64] estimate scene geometry by matching keypoints and triangulating points across multiple images. However, they require numerous images, precise matching, and high computational power, while struggling with occlusions, textureless regions, and scalability.

With the rise of neural rendering [53], data-driven techniques have transformed the field by enabling high-fidelity 3D reconstruction and novel view synthesis (NVS) with significantly fewer input images. Notably, NeRFs [33] introduced an implicit scene representation with coordinate-based Multi Layer Perceptrons (MLP) that can produce highly realistic renderings, but require large training times due to an expensive volumetric rendering process. More recently, 3DGS [21] has emerged as an efficient alternative, adopting an explicit representation with anisotropic 3D Gaussian primitives and tiled rasterization to achieve real-time rendering while maintaining high visual quality. These advancements mark a paradigm shift in 3D reconstruction, bridging the gap between traditional geometry-based methods and neural approaches. Building on 3DGS, [16] introduces an algorithm for mesh extraction and a regularization term to encourage 3D Gaussians to align with a surface and facilitate mesh extraction, while [18] simplifies 3D modeling by adopting flat 2D Gaussians, enabling faster rendering and reduced storage requirements. However, 3D and 2DGS models focus solely on scene appearance and geometry, lacking object-level understanding. [60] addresses this by lifting 2D semantic masks to 3D with identity-encoded Gaussians for instance grouping. Yet, their method relies on view-consistent masks obtained by a video object tracker, making it prone to failures with similar or intermittently oc-

[1] https://github.com/ksnxr/GWC_solution

cluded objects. In contrast, [27] employs 3D-aware mask association, matching projected Gaussians to 2D masks and assigning group IDs based on maximum overlap, leading to a better differentiation of similar objects. Similarly, [47] introduces an optimal solver for enhancing the accuracy and efficiency of embedding 2D semantic masks in 3DGS reconstructions.

Plant reconstruction. Capturing 3D information for plant phenotyping has traditionally relied on ranging sensors like LiDAR [25], RGB-D cameras like Intel RealSense [38], or RGB-based Structure-from-motion (SfM) and MVS [5, 11, 14, 35, 41, 51]. However, these methods are often costly and struggle to capture thin plant structures, leading to noisy and sparse point clouds. To address these challenges, [12] uses a robotic platform equipped with both LiDAR and camera sensors to reconstruct 3D plant structures from multiple sensing modalities. More recently, methods relying entirely on RGB images have emerged. For instance, [3, 17, 63, 65] evaluate several NeRF variants [7, 34, 52] for 3D reconstruction and NVS of various plants, including corn, tomatoes, and fruit trees, across different levels of complexity. However, these methods focus solely on 3D reconstruction without explicit plant phenotyping. To integrate plant trait analysis, PeanutNeRF [42] employs Nerfacto [52], a fast NeRF variant based on [34], for both 3D reconstruction and phenotypic analysis, extracting traits such as node count and flowering. However, it is limited to coarse plant structures and applies only to isolated plants in controlled environments with minimal occlusion. Similarly, [58] relies on NeRF for 3D reconstruction of rice panicles, combining YOLOv8 and SAM for instance segmentation and trait estimation, such as length and volume. However, their approach focuses on reconstructing single rice panicles one at a time, limiting its applicability in real-word scenarios. More scalable solutions are proposed by [32, 48], which extend NeRFs by mapping a 3D point not only to density and color but also to semantic information, successfully segmenting dozens of fruits in orchards and greenhouses.

3DGS has been proposed as a promising alternative to NeRF-based plant phenotyping [37, 46, 50]. However, these studies have so far remained exploratory at the single plant level, without providing semantic insights for plant trait analysis. In this work, we propose a mechanism to identify and segment wheat head instances in 3DGS reconstructions of wheat canopies in field conditions, thus representing the first work using radiance fields for 3D phenotypic trait extraction at scale.

3. Methodology

Our method (Fig. 2) is divided into four main sub-parts: wheat head segmentation on 2D images (Sec. 3.1), 3D reconstruction of wheat canopies (Sec. 3.2), 3D instance segmentation of individual wheat heads (Sec. 3.3 and Sec. 3.4), and phenotypic trait extraction (Sec. 3.5).

3.1. 2D Wheat Head Segmentation

We start with a collection of $\mathcal{C}$ unposed multi-view input images $\mathcal{I} = \{I_1, \dots, I_C\}$ of a wheat field. Detecting and precisely segmenting individual wheat heads from real-world images remains challenging due to their randomly scattered and densely-packed distribution. To address this, we adopt a two-step process that relies on a pre-trained YOLOv5 model [39] fine-tuned on the Global Wheat Head Dataset [10], which covers different locations, development stages, and capture conditions, followed by segmentation with SAM [23]. We utilize the pretrained YOLOv5 model given its publicly available weights specifically optimized for wheat head detection, and to demonstrate the robustness of our method. We use this model to generate a set of bounding boxes for each image I_i, $\mathcal{B}^i = \{b_1, \dots, b_{N_i}\}$, where N_i is the number of detected wheat heads on it, and provide them as prompts to SAM [23] individually, to obtain precise 2D segmentation masks for each detected wheat head. By iterating this procedure over all detections and images, we obtain a collection of 2D segmentation masks $\mathcal{M} = \{\mathcal{M}^1, \dots, \mathcal{M}^C\}$, where $\mathcal{M}^i = \{M_1, \dots, M_{N_i}\}$ is the set of single-wheat-head semantic masks associated with image I_i. Since SAM's segmentation output is generated independently for each image and is instance-agnostic, two distinct masks M from different images can correspond to the same physical wheat head in the canopy.

3.2. 3D Gaussian Splatting

We adopt 3DGS [21] as our scene representation for 3D reconstruction given its more straightforward editing capabilities compared to NeRF-based representations. Specifically, a 3D scene—one wheat plot in our case, as described in Sec. 4—is parameterized as a set of learnable 3D Gaussian primitives $\mathcal{G} = \{G_k\}_{k=1}^{K}$. Each Gaussian G_k is characterized by its centroid position $\mathbf{p}_k \in \mathbb{R}^3$, 3D scale vector $\mathbf{s}_k \in \mathbb{R}^3$, a quaternion $\mathbf{q}_k \in \mathbb{R}^4$ representing rotation, opacity $\alpha_k \in \mathbb{R}$, and color features $\mathbf{c}_k$ encoded by spherical harmonics (SH) coefficients. In its rendering process, 3DGS adopts a point-based rasterization approach that blends 3D Gaussians sorted by depth onto a 2D image plane using alpha compositing. A pixel-feature X is computed as:

$$X = \sum_{k \in \mathcal{K}} x_k \alpha_k \prod_{j=1}^{k-1} (1 - \alpha_j) = \sum_{k \in \mathcal{K}} x_k a_k T_k \qquad (1)$$

where the property x_k can be view-dependent color, depth, or other optimizable features of each Gaussian G_k, and T_k is the transmittance. The rendered image is compared to the ground truth camera view via photometric loss, which

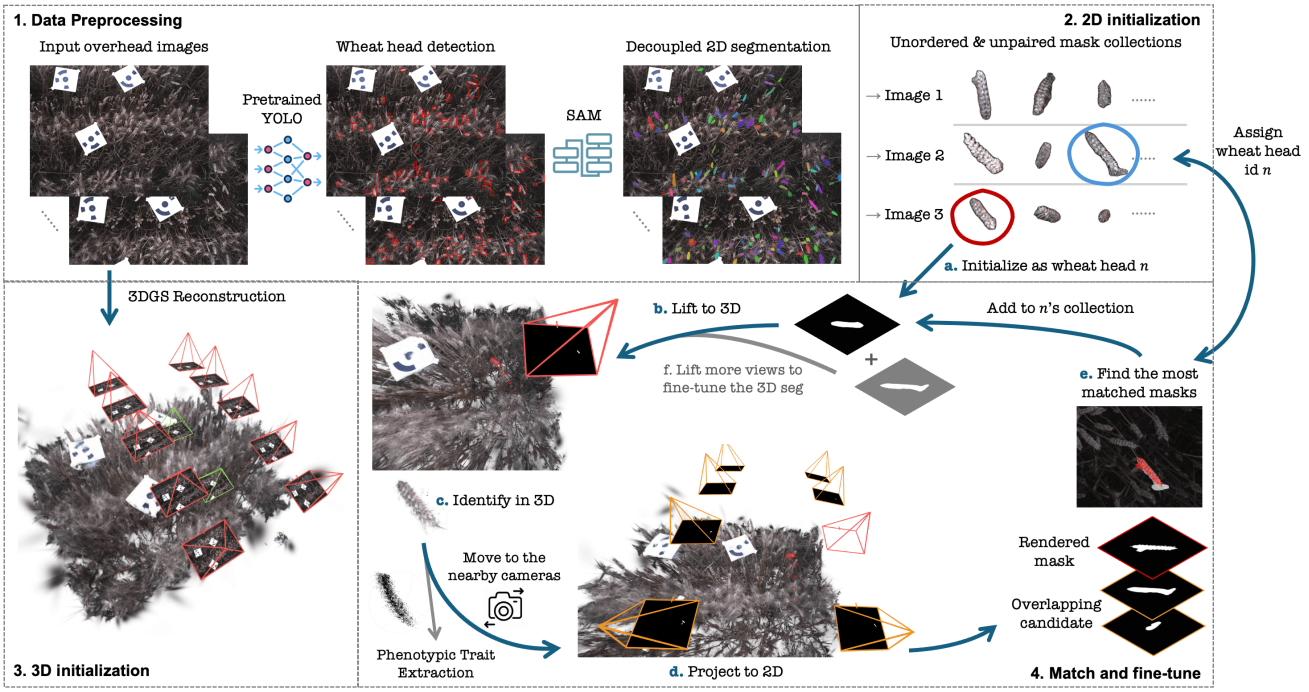


Figure 2. **Overview of our pipeline**: Given a set of RGB images capturing our target wheat field plot with a system of overhead cameras, we extract 2D segmentation masks of detected wheat heads (Sec. 3.1) and reconstruct a 3D representation of the plot using 3D Gaussian Splatting (Sec. 3.2) as initialization. For robust 3D wheat head segmentation (Sec. 3.3), we propose a **match-and-fine-tune** strategy (Sec. 3.4) that iteratively associates collections of decoupled masks and refines the 3D Gaussian representation of each segmented wheat head by alternating between lifting 2D masks to 3D and projecting 3D segmentations back to other views.

provides the learning signal to update the parameters of the Gaussians involved in rendering each pixel of the image.

3.3. 3D Wheat Head Segmentation

In this section we provide the preliminaries and the problem definition for 3D segmentation. Given a reconstructed 3DGS scene of a wheat field plot parameterized by 3D Gaussians $\mathcal{G}$ and containing a set of wheat heads $\mathcal{N}$, our goal is to identify the subset $\mathcal{G}_n \subset \mathcal{G}$ for each distinct wheat head $n \in \mathcal{N}$.

Suppose we have $\mathcal{L}$ 2D binary masks exclusively associated with one wheat head n, denoted as $\{M^\ell\}_n$ where $\ell \in \mathcal{L}$, $|\mathcal{L}| \leq |\mathcal{C}|$, and pixels with value 0 represent the background and 1 denotes the foreground (i.e. wheat head). This naturally leads to a 3D scene segmentation problem: assign a binary label $W_k \in \{0, 1\}$ to each 3D Gaussian G_k, indicating whether it corresponds to the targeted wheat head, by projecting the 2D binary masks M^ℓ into the 3D space.

In contrast to [61] that uses gradient descent to iteratively optimize a learnable embedding (from which the label assignment W_k can be derived) as a feature in Eq. 1, we adopt the method introduced in [47], which directly solves for the label W_k in closed form via integer linear programming. For each Gaussian G_k in a reconstructed scene $\mathcal{G}$, we set $x_k = W_k$ in Eq. 1 and optimize W_k while keeping all other attributes fixed. The 3D segmentation problem for a specific wheat head can then be formulated as solving for $\{W_k\}$ by minimizing the objective function $\mathcal{F}$ defined as the mean absolute error between the rendered 2D segmentation mask and the ground truth 2D mask:

$$\begin{aligned} \min_{\{W_k\}} \quad \mathcal{F} &= \sum_{\ell \in \mathcal{L}} \left| \mathcal{R}\left(\{G_k\}, \{W_k\}\right) - M^\ell \right| \\ &= \sum_{\ell \in \mathcal{L}} \sum_{p \in M^\ell} \left| \sum_{k \in \mathcal{K}} W_k \alpha_k T_k - M^\ell(p) \right| \end{aligned} \tag{2}$$

subject to $W_k \in \{0, 1\}$, where $\mathcal{R}$ is the differentiable rasterizer that renders each pixel value by blending $\mathcal{K}$ depth-sorted Gaussians, p represents a pixel in the provided binary mask M^ℓ, and $\ell \in \mathcal{L}$ specifies the view from which M^ℓ is available. We follow the approach in [47] to solve the optimal assignment for $\{W_k\}$ using majority vote across views and a background bias to account for noise in the masks.

Intuitively, the more ground truth masks M_n^ℓ are provided from different views ℓ for a wheat head n, the more accurately the subset of 3D Gaussians $\mathcal{G}_n = \{G_k \mid W_k = 1\}$ will represent the actual wheat head.

3.4. Multi-view Instance Association

The main challenge in our problem setup is the absence of associations between 2D wheat head masks detected across

different views, i.e. $\{M^\ell\}_n$, as defined in Sec. 3.3, is unknown.

Existing methods [61] either employ a video tracker [8] to propagate and associate 2D masks, which is ineffective in our case due to sparse viewpoints, and the densely-packed and repetitive structure of wheat canopies, or require handcrafted point prompts [47]. Thus, we developed a fully automatic iterative **match-and-fine-tune** strategy to address this challenge effectively for all wheat heads.

For a specific wheat head n, we are only certain that a single binary mask M_n^ℓ from one view ℓ is associated with it. By minimizing the discrepancy between the rendered ($\hat{M}_n^\ell$) and the provided mask (M_n^ℓ), we can optimize the binary label $W_k \in \{0, 1\}$ of each 3D Gaussian k as outlined in Sec. 3.3. However, intuitively, since only a single view ℓ is available, the estimated set $\{\hat{W}_k\}$ will be less accurate in representing the complete 3D structure of the wheat head when lifting the 2D segmentation to 3D.

To improve accuracy, we project the estimated $\{\hat{W}_k\}$ to another view $\ell' \in \mathcal{L}$, where $\ell' \neq \ell$, and render a binary mask $\hat{M}_n^{\ell'}$. Such rendered masks are often distorted due to insufficient views for accurate 3D segmentation. Hence, for each projected mask $\hat{M}_n^{\ell'}$ in a camera view, we identify the potential matching binary mask $M_{n'}^{\ell'}$ with the highest Intersection over Union (IoU) from the previously obtained 2D segmentation masks collection, which corresponds to a candidate matching wheat head n'. If the precision between $\hat{M}_n^{\ell'}$ and $M_{n'}^{\ell'}$ (we use precision because we observe $\hat{M}$ is often stretched) is larger than an empirically set threshold of 0.8, then we propose $n = n'$, that is, M_n^ℓ and $M_{n'}^{\ell'}$ correspond to the same wheat head. We continue this procedure for the remaining views in $\mathcal{L}$ and collect a new set of binary masks $\mathcal{M}_n = \{M_n^\ell, M_n^{\ell'_1}, M_n^{\ell'_2}, \dots\}$, representing the same matched wheat head n across different views. Note that we often have $|\mathcal{M}_n| < |\mathcal{L}|$ due to the limited camera coverage and the missed detection of wheat heads in 2D. We again solve for the linear optimization in Eq. 2, but now restrict the summation to include only masks $M^\ell \in \mathcal{M}_n$. A weighted majority vote approach, as introduced in [47], is used to resolve the contradiction within the mask set. The output assignment $\{\hat{W}_k\}$ now better identifies the 3D Gaussians belonging to wheat head n, that is, we have found a subset $\mathcal{G}_n = \{G_k \mid \hat{W}_k = 1\}$ that more accurately represents wheat head n from its matching detections across views. We then assign a unique wheat head ID $n > 0$ as an additional attribute to each Gaussian $G_k \in \mathcal{G}_n$. Finally, we exclude the matched masks $\mathcal{M}_n$ from the collection for further iterations, and repeat the process until there are no masks left to process.

3.5. 3D Phenotypic Trait Extraction

Once 3D point clouds of individual wheat head instances were obtained, they were subject to preprocessing and trait extraction steps. The preprocessing step was realized as follows: 1) random subsampling to 5000 points (if greater than 5000); 2) running HDBSCAN [29] to extract the dominant cluster of points—likely to correspond to a wheat head; 3) running robust Statistical Outlier Removal (SOR). Subsequently, we obtained per wheat head length, width, and volume. For each wheat head, length is extracted by projecting 3D points onto a plane spanning through the first and second principal components (1st-2nd-PC plane), fitting a 2D smoothing spline, and evaluating an approximation of the related arc length integral. Width is computed as the robust maximum distance (99th percentile) of points from the 1st-2nd-PC plane, and volume is computed from the convex hull obtained with the Quickhull algorithm [4]. All (hyper)parameters were chosen by trial and error to ensure generalizability across the used datasets. Further implementation details are given in the accompanying open-source code. The extracted traits are analyzed in Sec. 5.3.

4. Data

Setup. Data collection was performed on July 17 2024, and relied on a small-scale wheat phenotyping experiment. The setup comprised seven plots, each measuring approx. 1.5 m^2 and having six seeding rows, each row related to a different wheat genotype (see Fig. 1 for a plot example). A setup overview is presented in the suppl. material.

Images. Image acquisition was conducted using the cable-mounted camera rig system from ETH Zürich's Field Phenotyping Platform (FIP) [22]. We captured 36 images (12 MP) per plot from 12 identical cameras with 35 mm lenses. Three coded markers placed on each plot facilitated SfM, scale setting, and alignment with reference scans. SfM was performed in Agisoft Metashape (St. Petersburg, Russia) to obtain camera calibrations and sparse point clouds. Additionally, we generated dense MVS point clouds for comparison with our proposed workflow.

Laser scans. Reference measurements were obtained on the same day using a FARO Focus 3D S 120 (FARO Technologies, Inc, FL, USA) terrestrial laser scanner (TLS) with full resolution (1.6 mm @ 10 m) and the highest quality setting. The scanning setup comprised 19 scans at a few meters distance, aside and above the canopy using a tripod and a custom mount (see suppl. material). The scans were registered in the FARO SCENE 2022.1.0 software using a target-based algorithm, relying on six laser scanning reference spheres placed within the scene, and achieving a mean alignment error of 3 mm. This was followed by a coarse alignment with the dense MVS point cloud using the Kabsch algorithm [20] and corresponding marker points identified in both point clouds. Finally, the registered scans

were precisely aligned with 3DGS centroids of all wheat heads (obtained following Sec. 3.4) using the ICP algorithm [6] on the subsampled point clouds. The alignment of the corresponding scene elements between the TLS and 3DGS datasets was assessed to be within 10 mm on average based on visual inspection. As scanning took approximately 6 h, the internal geometry of the scene could change during the acquisition due to mild wind gusts and plant motion, which affects the quality of the scan registration [30]. Hence, even though TLS is commonly considered as "ground truth" for built and urban environments, in this challenging scenario, it should be considered as an independent control of comparable reconstruction quality.

5. Results

In this section, we present results in terms of NVS (Sec. 5.1), wheat head detection and segmentation by comparing to state-of-the-art models (Sec. 5.2), and geometric validation against TLS data (Sec. 5.3).

5.1. Novel view synthesis

We evaluated different differentiable rendering methods on our dataset to determine the optimal underlying scene representation for 3D segmentation. In order to robustly assess the reconstruction quality, we used 30 images for training and withheld 6 images for evaluation. For all plots, we selected the test views such that they were out-of-the-distribution compared to the train views (see Fig. 1).

We report Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) in Tab. 1, which are standard metrics for quantitative evaluation of visual reconstruction quality of 3D scenes [3]. 3DGS (based on gsplat [62] implementation) achieves the best results on image quality metrics, followed by Nerfacto, with a considerable margin in SSIM and LPIPS. In Fig. 3 we show qualitative comparison, highlighting the greater level of detail achieved by 3DGS, especially on fine-grained structures such as wheat head awns. Results for additional baselines can be found in the suppl. material. We note that other baselines based on the original 3DGS codebase present lower results in pixel-wise metrics, but achieve better results in perceptual metrics like LPIPS. We believe the issue lies in an image transformation problem causing a translation shift by a few pixels for which we could not find a solution. This problem only affects 2D evaluations and does not cause visible quality degradation in the 3D reconstructions. We used reconstructions from the original 3DGS method in the rest of our pipeline for ease of integration with our 3D instance segmentation solution.

5.2. Wheat head segmentation

For obtaining 2D input segmentation masks, we experimented with a combination of a pretrained YOLO & SAM; and a zero-shot approach based on Grounded SAM [40]. The Grounded SAM approach used a "wheat head" prompt and incorporated SAHI (Slicing Aided Hyper Inference) [1], which is designed to improve detections of multiple small objects. The results of the predictions by these two methods can be seen in Fig. 4b and Fig. 4c respectively. More wheat heads are detected with the Grounded SAM approach, however, the segmentations are more noisy. Favoring reliability, we used the pretrained YOLO & SAM approach in our automated 3D segmentation solution despite the lower amount of detections, given that this limitation can be effectively alleviated with detections from other views.

To evaluate 3D wheat head segmentation methods on our data we annotated the bounding boxes for each observable wheat head on a randomly-chosen test view per plot. The wheat heads instances were then segmented by giving these bounding boxes as input prompts to SAM. We visually verified the quality of the output masks.

We compared our method with FruitNeRF [32], providing the same input segmentation masks. Quantitative results of both methods compared to the ground truth are presented in Tab. 2. Additionally, qualitative comparisons of our method and FruitNeRF to the ground truth mask of one plot are illustrated in Fig. 4d and Fig. 4e. Note that many more wheat heads are detected than in the input masks,

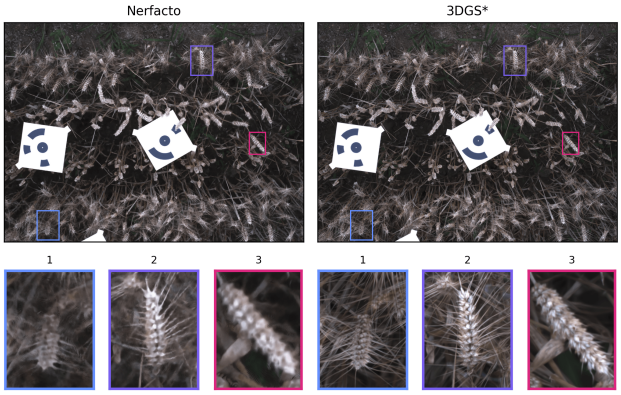


Figure 3. Comparison of Nerfacto and 3DGS* (gsplat implementation) renderings from a test view. Matching zoom regions (1-3) below each image highlight structural details.

Table 1. Quantitative comparison for NVS on our test set. We evaluate radiance fields methods (after 30k iterations) based on image quality metrics, average training time, and storage of the trained model. Colors highlight best, second-best and third-best method on each metric. 3DGS*: gsplat implementation of 3DGS.

Method	SSIM↑	PSNR↑	LPIPS↓	Time (min)	Storage (GB)
Instant-NGP [34]	0.662	20.891	0.506	39	0.185
Nerfacto [52]	0.769	25.387	0.384	45	0.164
FruitNeRF [32]	0.752	23.382	0.422	47	0.236
3DGS* [62]	0.843	25.447	0.226	146	0.557

Table 2. Quantitative results against ground truth 2D segmentation masks. Best results per metric are highlighted in red.

Method	IoU (%)	Precision (%)	Recall (%)	F1	MSE	SSIM
FruitNeRF [32]	0.34	0.95	0.35	0.50	0.05	0.70
Ours	0.50	0.81	0.57	0.67	0.06	0.90

highlighting the effective incorporation of multi-view detections into our 3D representation.

5.3. Geometric validation and applications

We validated the geometric accuracy of our extracted wheat heads by comparing them to aligned TLS data of individual wheat head instances in terms of their 3D morphological traits (length - L, width - W, volume - V), following Sec. 3.5. We compared the Gaussian centers (i.e. point clouds) of our 3DGS results to TLS and MVS point clouds. However, only the point clouds from our 3DGS-based pipeline were segmented by instance of individual wheat heads. To enable comparison across modalities, we: 1) took single georeferenced 3DGS wheat heads and filtered out all TLS and MVS data that were >15 mm away from the closest 3DGS point; 2) assigned oriented bounding boxes to each 3DGS wheat head instance, applying a buffer of 10 mm (accounting for the alignment uncertainty), and extracted the matching wheat head instances from TLS and MVS data (Fig. 5).

We compared 3DGS- and MVS-based estimates to TLS on a per-instance and a per-row-average (genotype) basis by linear regression after outlier removal. The per-row-average analysis investigates the potential for phenotyping applications. The results are summarized in Tab. 3. The observed significant, but moderate correlations (ρ) indicate general agreement between the datasets, however, with a high per-instance noise (see e.g. mean absolute percentage error (MAPE) or mean absolute error (MAE)). Computing per-row averages notably decreases MAPE in all cases, increases ρ for L and W, but not for V. These results indicate that: 1) 3DGS and MVS perform comparably well; 2) noise is too high for confident per-instance phenotyping; 3) averaging reduces noise levels to the point that phenotypic data can be used to distinguish different genotypes with moderate to high confidence; 4) the extracted L and W values represent the same real-world physical quantities. Yet, the extracted V values are either too noisy, or image-based and scanning-based estimates capture different aspects of plant structure.

To further test the hypothesis that the extracted traits could be useful for phenotyping applications (i.e. to measure statistically significant differences between genotypes) we conducted a one-way analysis of variance (ANOVA) for each measurement method and present the results in Table Tab. 4. The derived F-statistics were significantly (P-value $\ll 0.01$) and notably higher than 1, indicating rejection of the null-hypothesis (no significant difference between genotype means) and strong discriminative power of the derived quantities. For L, TLS provided the largest between-genotypes variability. However, for W and V, 3DGS was notably more discriminative, hinting stronger usability in phenotyping applications than the reference TLS data.

Table 3. Per-instance and per-row-average agreement: TLS (reference) vs. 3DGS and MVS. We report correlation (ρ), mean absolute error (MAE), and mean absolute percentage error (MAPE) for length (L), width (W), volume (V). MAE units are in cm for L and W, and cm^3 for V. P-value $\ll 0.01$ in each per-instance case, ≤ 0.05 in each per-row-average case, except 3DGS-V. Best results per trait and metric are highlighted in red.

		per-instance			per-row-average		
		L	W	V	L	W	V
ρ	MVS	0.51	0.35	0.40	0.74	0.53	0.32
	3DGS	0.51	0.27	0.32	0.69	0.43	0.05
MAE	MVS	1.51	0.35	12.57	0.58	0.19	9.64
	3DGS	1.48	0.25	10.72	0.79	0.13	6.12
MAPE	MVS	16.0	26.0	47.2	5.9	15.0	39.9
	3DGS	15.1	18.3	40.2	8.1	9.9	24.4

Table 4. One-way ANOVA F-statistics for length (L), width (W), and volume (V) based on 2389 samples of 42 populations (P-value $\ll 0.01$ in each case). Best results per trait are highlighted in red.

L			W			V		
TLS	3DGS	MVS	TLS	3DGS	MVS	TLS	3DGS	MVS
15.2	11.2	10.1	5.2	35.0	8.5	6.9	10.8	7.1

6. Discussion

We showed that the combination of 3DGS with our multi-view instance segmentation pipeline effectively captures the complex geometry of wheat canopies allowing to extract, count and measure hundreds of individual wheat heads in 3D. Unlike previous RGB image-based approaches for 3D reconstruction that typically require dense spatial coverage of viewpoints, our method achieved high-quality reconstructions with only 30 views per plot and with a limited viewpoint distribution (Fig. 1). This is particularly important for practical field applications, where capturing dense and diverse viewpoints may be infeasible or time-consuming. Comparatively, NeRF-based methods like FruitNeRF [32] tackled simpler scenes and tasks (fruit counting in horticulture) using several hundreds of images with good spatial coverage. Meanwhile, [13] proposed an MVS-based approach to measure fruits from 2D images with precise amodal segmentation masks and depth maps, but cannot perform volume estimation.

Beyond good 3D reconstruction and segmentation, we show that our method allows for in-field morphological trait extraction with a sufficient precision for distinguishing between different genotypes, facilitating phenotyping applications in a breeding context. So far, similar achieve-



(a) Pretrained YOLO (b) Pretrained YOLO + SAM (c) Grounding DINO + SAHI (d) GT vs. FruitNeRF (e) GT vs. Ours

Figure 4. **Left: visualization of 2D detection and segmentation** of wheat heads on a train image from (a) pre-trained YOLOv5, (b) Segment Anything with detected bounding boxes as prompt, and (c) state-of-the-art GroundedSAM2 version which combines Grounding DINO 1.5 with SAHI (Slicing Aided Hyper Inference), with "wheat head" as text prompt. **Right: qualitative evaluation of novel view mask rendering** from the 3D segmentation obtained by our pipeline (using (b)). (d) and (e) compare the projection of 3D instance segmentation onto 2D in a novel view with human-labeled wheat head segmentation. Green represents the overlap between rendered masks and ground truth (GT), i.e. correct segmentation of wheat head; orange indicates false positive segmentation; and red represents wheat heads not identified in the 3D instance segmentation, resulting in their absence in the projected 2D masks.

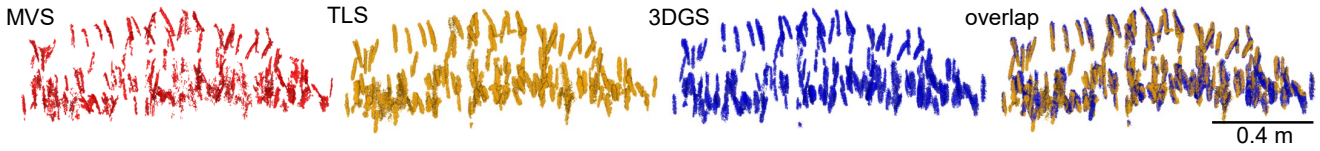


Figure 5. Corresponding wheat head instances of one experimental plot extracted from all three datasets.

ments have been demonstrated only using expensive high-end laser scanning instruments [26, 55], attaining comparable results with lower uncertainty (MAPE per-instance: L of 4-5% and W of 12-32%; MAPE per-genotype-average: L of 15% and W of 24%).

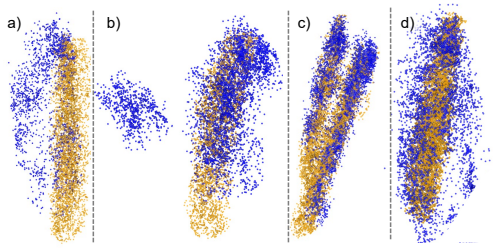


Figure 6. Reoccurring failure cases causing disagreement between the reference TLS (orange) and 3DGS (blue): a) incorrect reconstruction at lower canopy levels and scene edges; b) and c) incorrect instance segmentation; d) ambigous geometry reconstruction due to strongly expressed awns.

Despite reasonably promising results, we observed some disagreements between 3DGS and the reference TLS. There are several common failure cases (Fig. 6): a) splats of wheat heads at the lower canopy levels and on the scene edges (limited number of views) can diverge from the real wheat head surface and can fail to reconstruct the bottom part of the wheat head; b) occasional errors in instance segmentation lead to multiple splat clusters being related to a single instance—sometimes capturing wheat head-unrelated foliage, or c) merging multiple wheat heads together; d) splats partially capturing structure of strongly expressed awns (spikes) for some wheat varieties, but insufficiently well for their full reconstruction. The latter phenomena is a likely cause for the observed strong disparity between 3DGS- and TLS-based volume estimates, as the TLS is inherently unable to capture such fine structural details due to finite laser beam footprint size (Tab. 3). Future work directions to address these issues may include introducing prior knowledge about wheat heads shape in a similar way to [28, 32], and ensuring robustness to environmental disturbances such as wind.

7. Conclusion

In this work, we address 3D reconstruction of wheat canopies and instance segmentation of wheat heads from multi-view images, using 3D Gaussian Splatting (3DGS), a pretrained wheat head detector, and the Segment Anything Model (SAM). We handle instance segmentation iteratively, by annotating Gaussians using maximal information available from inconsistent binary segmentation masks across views. Our approach demonstrates the effectiveness of 3DGS for the 3D reconstruction of wheat canopies and instance segmentation of wheat heads in field conditions. Comparisons against state-of-the-art NeRF-based methods for this task highlight superior reconstruction quality and segmentation performance of our approach. Furthermore, evaluations against terrestrial laser scan data demonstrate that our method achieves sufficient accuracy for high-throughput field phenotyping of wheat head morphological traits, including length, width, and volume.

References

[1] Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing aided hyper inference and fine-tuning for small object detection. *2022 IEEE International Conference on Image Processing (ICIP)*, pages 966–970, 2022. 6

[2] José Luis Araus, Shawn C. Kefauver, Mainassara Zaman-Allah, Mike S. Olsen, and Jill E. Cairns. Translating High-Throughput Phenotyping into Genetic Gain. *Trends in Plant Science*, 23(5):451–466, 2018. 1

[3] M. A. Arshad, T. Jubery, J. Afful, A. Jignasu, A. Balu, B. Ganapathysubramanian, S. Sarkar, and A. Krishnamurthy. Evaluating Neural Radiance Fields (NeRFs) for 3D Plant Geometry Reconstruction in Field Conditions. *Plant Phenomics*, 6, 2024. 1, 3, 6

[4] C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.*, 22(4):469–483, 1996. 5

[5] Juliane Bendig, Andreas Bolten, and Georg Bareth. Uav-based imaging for multi-temporal, very high resolution crop surface models to monitor crop growth variability. *Photogrammetrie - Fernerkundung - Geoinformation*, 2013(6): 551–562, 2013. 3

[6] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 6

[7] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su. TensoRF: Tensorial Radiance Fields. In *ECCV*, 2022. 3

[8] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1316–1326, 2023. 5

[9] Daohan Cui, Pengfei Liu, Yunong Liu, Zhenqing Zhao, and Jiang Feng. Automated Phenotypic Analysis of Mature Soybean Using Multi-View Stereo 3D Reconstruction and Point Cloud Segmentation. *Agriculture*, 2025. 1

[10] Etienne David, Mario Serouart, Daniel Smith, Simon Madec, Kaaviya Velumani, Shouyang Liu, Xu Wang, Francisco Pinto, Shahameh Shafiee, Izzat S. A. Tahir, and et al. Global wheat head detection 2021: An improved dataset for benchmarking wheat head detection methods. *Plant Phenomics*, 41, 2021. 2, 3

[11] I. Drofova, H. Wang, W. Guo, M. Pospisilik, M. Adamek, and J. Valouch. 3D reconstruction of a group of plants by the ground multi-image photogrammetry method. In *International Conference Radioelektronika (RADIOELEKTRONIKA)*, pages 1–4, 2023. 3

[12] F. Esser, R. A. Rosu, A. Cornelißen, L. Klingbeil, H. Kuhlmann, and S. Behnke. Field Robot for High-Throughput and High-Resolution 3D Plant Phenotyping: Towards Efficient and Sustainable Crop Production. *IEEE Robotics & Automation Magazine*, 30(4):20–29, 2023. 3

[13] Jordi Gené-Mola, Mar Ferrer-Ferrer, Eduard Gregorio, Pieter M. Blok, Jochen Hemming, Josep-Ramon Morros, Joan R. Rosell-Polo, Verónica Vilaplana, and Javier Ruiz-Hidalgo. Looking behind occlusions: A study on amodal segmentation for robust on-tree apple fruit size estimation. *Computers and Electronics in Agriculture*, 209:107854, 2023. 7

[14] Jeffrey K. Gillan, Jason W. Karl, Michael Duniway, and Ahmed Elaksher. Modeling vegetation heights from high resolution stereo aerial photography: An application for broad-scale rangeland monitoring. *Journal of Environmental Management*, 144:226–235, 2014. 3

[15] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M. Seitz. Multi-View Stereo for Community Photo Collections. In *ICCV*, 2007. 2

[16] A. Guédon and V. Lepetit. SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering. In *CVPR*, 2024. 2

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

[18] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. 1, 2

[19] Andreas Hund, Lukas Kronenberg, Jonas Anderegg, Kang Yu, and Achim Walter. Non-invasive field phenotyping of cereal development. In *Advances in breeding techniques for cereal crops*, page 249–292. Burleigh Dodds Science Publishing, 2019. 1

[20] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Foundations of Crystallography*, 32(5): 922–923, 1976. 5

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 3

[22] Norbert Kirchgessner, Frank Liebisch, Kang Yu, Johannes Pfeifer, Michael Friedli, Andreas Hund, and Achim Walter. The ETH field phenotyping platform FIP: A cable-suspended multi-sensor system. *Functional Plant Biology*, 44:154–168, 2017. 5

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3

[24] Maria Klodt and Daniel Cremers. High-Resolution Plant Shape Measurements from Multi-view Stereo Reconstruction. In *ECCV 2014 Workshops*, 2014. 1

[25] Lukas Kronenberg, Steven Yates, Martin P Boer, Norbert Kirchgessner, Achim Walter, and Andreas Hund. Temperature response of wheat affects final height and the timing of stem elongation under field conditions. *Journal of Experimental Botany*, 72(2):700–717, 2020. 3

[26] Zhonghua Liu, Shichao Jin, Xiaoqiang Liu, Qiuli Yang, Qing Li, Jingrong Zang, Zhaofeng Li, Tianyu Hu, Zifeng Guo, Jin Wu, et al. Extraction of wheat spike phenotypes from field-collected lidar data and exploration of their relation-

ships with wheat yield. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 1, 8

[27] W. Lyu, X. Li, A. Kundu, Y.-H. Tsai, and M.-H. Yang. Gaga: Group Any Gaussians via 3D-aware Memory Bank. *arXiv:2404.07977 [cs]*, 2024. arXiv: 2404.07977. 3

[28] Federico Magistri, Thomas Läbe, Elias Marks, Sumanth Nagulavancha, Yue Pan, Claus Smitt, Lasse Klingbeil, Michael Halstead, Heiner Kuhlmann, Chris McCool, Jens Behley, and Cyrill Stachniss. A Dataset and Benchmark for Shape Completion of Fruits for Agricultural Robotics. *arXiv:2407.13304 [cs]*, 2024. arXiv: 2407.13304. 8

[29] Leland McInnes, John Healy, Steve Astels, et al. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017. 5

[30] Tomislav Medic, Jonas Bömer, and Stefan Paulus. 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, 2023. 6

[31] Tomislav Medic, Nicole Manser, Norbert Kirchgessner, and Lukas Roth. Towards wheat yield estimation in plant breeding from inhomogeneous lidar point clouds using stochastic features. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:741–747, 2023. 1

[32] Lukas Meyer, Andreas Gilson, Ute Schmid, and Marc Stamminger. FruitNeRF: A unified neural radiance field based fruit counting framework. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8, 2024. 3, 6, 7, 8

[33] 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. *Commun. ACM*, 65(1): 99–106, 2021. 1, 2

[34] T. Müller, A. Evans, C. Schied, and A. Keller. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 1, 3, 6

[35] Toshifumi Murakami, Mamiko Yui, and Koichi Amaha. Canopy height measurement by photogrammetric analysis of aerial images: Application to buckwheat (Fagopyrum esculentum Moench) lodging evaluation. *Computers and Electronics in Agriculture*, 89:70–75, 2012. 3

[36] Thuy Tuong Nguyen, David C. Slaughter, Julin N. Maloof, and Neelima Sinha. Plant phenotyping using multi-view stereo vision with structured lights. *Proceedings of SPIE*, 9866, 2016. 1

[37] Tommy Ojo, Thai La, Andrew Morton, and Ian Stavness. Splanting: 3D plant capture with gaussian splatting. In *SIGGRAPH Asia 2024 Technical Communications*, pages 1–4, Tokyo Japan, 2024. ACM. 2, 3

[38] Y. Pan, F. Magistri, T. Läbe, E. Marks, C. Smitt, C. McCool, J. Behley, and C. Stachniss. Panoptic Mapping with Fruit Completion and Pose Estimation for Horticultural Robots. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4226–4233, Detroit, MI, USA, 2023. IEEE. 3

[39] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[40] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: Assembling open-world models for diverse visual tasks, 2024. 6

[41] Lukas Roth and Bernhard Streit. Predicting cover crop biomass by lightweight UAS-based RGB and NIR photography: an applied photogrammetric approach. *Precision Agriculture*, 19(1):93–114, 2018. 3

[42] Farah Saeed, Jin Sun, Peggy Ozias-Akins, Ye Chu, and Changying Li. PeanutNeRF: 3D Radiance Field for Peanuts. In *CVPRW*, pages 6254–6263, Vancouver, BC, Canada, 2023. IEEE. 2, 3

[43] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 2

[44] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, 2016. 2

[45] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, 2006. 2

[46] Peng Shen, Xueyao Jing, Wenzhe Deng, Hanyue Jia, and Tingting Wu. PlantGaussian: Exploring 3D Gaussian splatting for cross-time, cross-scene, and realistic 3D plant visualization and beyond. *The Crop Journal*, page S2214514125000261, 2025. 3

[47] Q. Shen, X. Yang, and X. Wang. FlashSplat: 2D to 3D Gaussian Splatting Segmentation Solved Optimally. In *ECCV*, page 456–472, Berlin, Heidelberg, 2024. Springer-Verlag. 2, 3, 4, 5

[48] C. Smitt, M. Halstead, P. Zimmer, T. Läbe, E. Guclu, C. Stachniss, and C. McCool. PAg-NeRF: Towards Fast and Efficient End-to-End Panoptic 3D Representations for Agricultural Robotics. *IEEE Robotics and Automation Letters*, 9 (1):907–914, 2024. 3

[49] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Transactions on Graphics*, 25:835–846, 2006. 2

[50] Lewis A G Stuart, Darren M Wells, Jonathan A Atkinson, Simon Castle-Green, Jack Walker, and Michael P Pound. High-fidelity wheat plant reconstruction using 3D Gaussian splatting and neural radiance fields. *GigaScience*, 14: giaf022, 2025. 3

[51] P. Sunvittayakul, P. Kittipadakul, P. Wonnapinij, P. Chanchay, P. Wannitikul, S. Sathitnaitham, P. Phanthanong, K. Changwitchukarn, A. Suttangkakul, H. Ceballos, and S. Vuttipongchaikij. Cassava root crown phenotyping using three-dimension (3d) multi-view stereo reconstruction. *Scientific Reports*, 2022. 3

[52] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David

McAllister, and Angjoo Kanazawa. Nerfstudio: A Modular Framework for Neural Radiance Field Development. In *ACM SIGGRAPH 2023 Conference Proceedings*, New York, NY, USA, 2023. Association for Computing Machinery. 1, 3, 6

[53] A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi, T. Simon, C. Theobalt, M. Nießner, J. T. Barron, G. Wetzstein, M. Zollhöfer, and V. Golyanik. Advances in Neural Rendering. *Computer Graphics Forum*, 41(2):703–735, 2022. 2

[54] Bill Triggs, Philip F McLauchlan, Richard I Hartley, , and Andrew W Fitzgibbon. Bundle Adjustment — A Modern Synthesis. In *Vision Algorithms: Theory and Practice*, 2000. 2

[55] Fuli Wang, Fengping Li, Vishwanathan Mohan, Richard Dudley, Dongbing Gu, and Ruth Bryant. An unsupervised automatic measurement of wheat spike dimensions in dense 3d point clouds for field application. *Biosystems Engineering*, 223:103–114, 2022. 1, 8

[56] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *CVPR*, 2021. 2

[57] Sheng Wu, Yongjian Wang Weiliang Wen, Jiangchuan Fan, Chuanyu Wang, Wenbo Gou, and Xinyu Guo. MVS-Pheno: A Portable and Low-Cost Phenotyping Platform for Maize Shoots Using Multiview Stereo 3D Reconstruction. *Plant Phenomics*, 2020. 1

[58] Xin Yang, Xuqi Lu, Pengyao Xie, Ziyue Guo, Hui Fang, Haowei Fu, Xiaochun Hu, Zhenbiao Sun, and Haiyan Cen. PanicleNeRF: Low-Cost, High-Precision In-Field Phenotyping of Rice Panicles with Smartphone. *Plant Phenomics*, 6: 0279, 2024. 2, 3

[59] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 2020. 2

[60] M. Ye, M. Danelljan, F. Yu, and L. Ke. Gaussian Grouping: Segment and Edit Anything in 3D Scenes. In *ECCV*, 2023. 2

[61] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian Grouping: Segment and Edit Anything in 3D Scenes. *arXiv:2312.00732 [cs]*, 2023. arXiv: 2312.00732. 4, 5

[62] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for gaussian splatting, 2024. arXiv: 2409.06765. 6

[63] Albert J. Zhai, Xinlei Wang, Kaiyuan Li, Zhao Jiang, Junxiong Zhou, Sheng Wang, Zhenong Jin, Kaiyu Guan, and Shenlong Wang. CropCraft: Inverse Procedural Modeling for 3D Reconstruction of Crop Plants. In *arXiv:2411.09693v1*, 2024. 3

[64] Jingyang Zhang, Shiwei Li, Zixin Luo, Tian Fang, , and Yao Yao. Vis-mvsnet: Visibility-aware multi-view stereo network. *IJCV*, 131:199–214, 2023. 2

[65] J. Zhang, X. Wang, X. Ni, F. Dong, L. Tang, J. Sun, and Y. Wang. Neural radiance fields for multi-scale constraint-free 3D reconstruction and rendering in orchard scenes. *Computers and Electronics in Agriculture*, 217:108629, 2024. 3