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

Supplementary Material

A. Dataset setup details

We present an overview of our data collection setup on seven wheat plots in Fig. S1. Each wheat plot contained six rows of different wheat varieties. Image acquisition was performed using the Field Phenotyping Platform (FIP) of ETH Zürich [22]. The platform consists of a multi-view camera rig mounted on a SpiderCam (Spidercam robotics GmbH, Feistritz, Austria) cable system and is equipped with 13 cameras (Fig. S2). We used only 12 cameras (DFK 38UX304, 12 MP, The Imaging Source, Bremen, Germany) for their identical lens specifications (V3522-MPZ, 35 mm, The Imaging Source, Bremen, Germany) and field of view (FOV). For each of the seven plots, we captured three sets of 12 images, with an approximately 25 cm collinear shift in rig position between the sets, resulting in 36 views per plot for 3D reconstruction. Three coded ring markers were placed on each plot to aid Structure from Motion (SfM), set the scale, and enable alignment of reference laser scans. Marker coordinates were measured with a Trimble R10 GNSS device in RTK mode (1-2 cm positioning accuracy). SfM was performed in Agisoft Metashape (St. Petersburg, Russia) using all 36 images to obtain camera calibrations and a sparse point cloud per plot. This provided a basic input for our proposed workflow. In addition, the MVS pipeline was performed to obtain dense point clouds for comparing the proposed workflow against the traditional 3D photogrammetry reconstruction. Finally, Fig. S3 shows the custom laser scanner mount used in this study in addition to the tripod shown in Fig. S1.

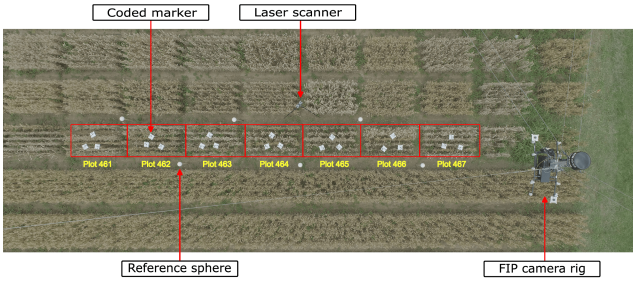

Figure S1. Overview of our data collection setup.

B. Additional NVS baselines

While gsplat [62] implementation of 3DGS outperforms other radiance field methods in terms of NVS, we also experimented with the original 3DGS implementation by Inria [21], as well as 2D Gaussian Splatting (2DGS) [18] and SuGaR [16] (Tab. S1). Notably, the latter two methods provide advantages for 3D mesh extraction, which may be desirable in certain workflows. As described in Sec. 5.1, we observed a pixel misalignment between the rendered evaluation views and ground truth views when using the original implementation of 3DGS and its variants, 2DGS and SuGaR. Such positional shifts significantly degrade pixel-wise image quality metrics, such as SSIM and PSNR, causing it to perform worse than NeRF-based methods, despite achieving higher perceptual image quality metrics like LPIPS. Regarding NeRF-based models, although FruitNeRF-big [32] has greater layer depth, hidden dimension, and overall capacity for modeling density and appearance, its performance degrades compared to the smaller version. We suspect the reason is that our limited amount of training views (30) compared to the original dataset the model was developed on leads to severe overfitting [32].

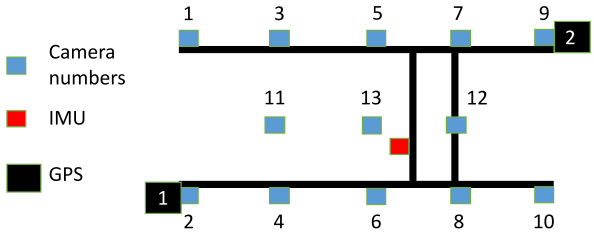

Figure S2. Schematic top view of the FIP multi-camera rig system.

Figure S3. Our custom mount for upside-down laser scans.

Table S1. **Quantitative comparison for Novel View Synthesis** on our dataset. We evaluate neural rendering methods based on image quality metrics, average training time, and stored model size. 3DGS*: gsplat implementation of 3DGS. Note that pixel-wise metrics (SSIM, PSNR) for 3DGS [21] and its variants, 2DGS [18] and SuGaR [16], are negatively affected by pixel misalignment between rendered and ground truth views due to a bug in data transformation, which does not severely affect patch-based metrics (LPIPS). The two types of SuGaR (coarse and refined) correspond to the sets of 3D Gaussians extracted at different stages of SuGaR's optimization for surface alignment.

Type	Method	SSIM↑	PSNR↑	LPIPS↓	Time (min)	Storage (GB)
NeRF-based	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
	FruitNeRF big	0.500	15.663	0.666	440	0.792
Gaussian-based	3DGS* [62]	0.843	25.447	0.226	146	0.557
	3DGS 7k iters [21]	0.651	20.549	0.333	31	0.996
	3DGS 15k iters	0.639	20.416	0.323	74	1.286
	2DGS [18]	0.560	20.593	0.241	72	-
	SuGaR coarse [16]	0.569	20.716	0.278	40	0.102
	SuGaR refined	0.549	20.520	0.290	84	0.488

C. Additional qualitative results

We provide additional qualitative results in Fig. S4, which demonstrate that 3DGS* produces renderings with fewer deviations from the ground truth image compared to Nerfacto, as evidenced by the reduced structural details visible in the difference maps.

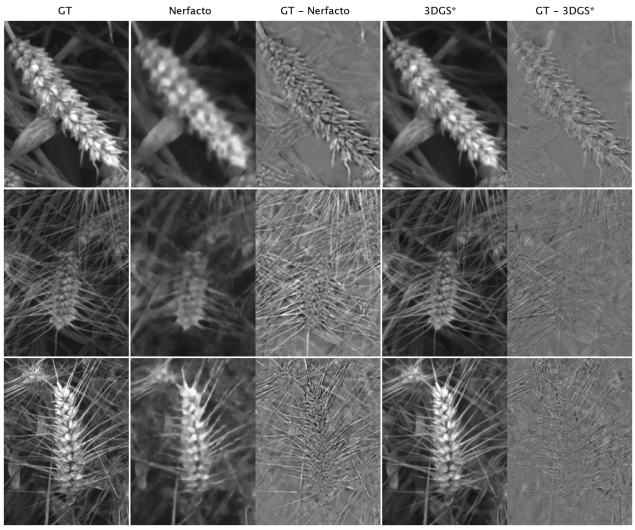

Figure S4. Comparison of wheat head renderings (same ones as in Fig. 3) to the ground truth image. From left to right: ground truth (GT), Nerfacto, GT-Nerfacto difference, 3DGS* (gsplat implementation), and GT-3DGS* difference. All difference maps are shown in identical grayscale range.

D. Additional laser scan comparisons

We repeated the comparison of the 3DGS, TLS and MVS-based wheat head length (L), width (W) and volume (V) estimates, per-instance and per-row average (genotype), after removing obvious 3D reconstruction failure cases as discussed in Sec. 6.

Failure cases were automatically detected as out-of-the-distribution samples of 2D L, W, and V values distributions; where the first and second dimensions of the respective 2D distributions were defined as TLS-based and 3DGS-based trait estimate values. The expected (failure-case-free) theoretical data distributions were determined by robustly fitting 2D Gaussians (by Minimum Covariance Determinant - MCD estimator) and detecting and removing all points that were outside the confidence interval (comparing squared mahalanobis distance with threshold value drawn from the Chi-squared distribution, 95th percentile, 2 degrees of freedom).

The updated results with mainly improved metrics are presented in Tab. S2. Eliminating the most prominent of these failure cases leads to notable increases in similarity between 3DGS-based and TLS-based estimates. MAE decreases from 1.48 to 0.73 cm, 0.25 to 0.21 cm, and 10.72 to 7.25 cm^3 for the per-instance comparison case for L, W and V respectively; and changes from 0.79 to 0.52 cm, 0.13 to 0.11 cm, and 6.12 to 4.48 cm^3 for the per-row-average case (on average MAE decreases 30%).

Table S2. Per-instance and per-row-average agreement after filtering out the failure cases: 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 MVS-V. Best results per trait and metric are highlighted in red.

		per-instance			per-row-average		
		L	W	V	L	W	V
ρ	MVS	0.55	0.35	0.36	0.75	0.55	0.31
	3DGS	0.78	0.33	0.39	0.73	0.58	0.39
MAE	MVS	1.09	0.31	10.00	0.51	0.18	8.42
	3DGS	0.73	0.21	7.25	0.52	0.11	4.48
MAPE	MVS	12.3	24.1	43.9	5.5	14.38	38.92
	3DGS	8.2	16.7	32.15	5.6	8.88	20.51