3D Reconstruction and Knowledge Distillation to Improve Multi-View Image Models to Explore Spike Volume Estimation in Wheat

Supplementary Material

A. Dataset and Acquisition

Wheat spikes were collected from a trait calibration panel. This panel consisted of i) historic varieties from Switzerland, France, and Germany with important post-green revolution varieties, ii) a diverse set of varieties widely tested in the framework of the EU projects INVITE ([73, 74]), and iii) varieties included in variety registration in Switzerland. To collect a diverse dataset, spikes of 83 and 82 different genotypes were imaged by RGB cameras (Fig. S1a) and sampled in 2023 and 2024, respectively, with 72 overlapping genotypes. Tagged spikes (Fig. S1b) were imaged and sampled three and two times in each season: (i) at flowering (June 9, 2023, and June 12, 2024); (ii) between flowering and maturity (June 29, 2023, and July 5, 2024); and (iii) at maturity (July 11, 2023). The mean spike volume across both years was 4649.06 mm^3, with a mean standard deviation of 1234.26 mm^3. The mean ($\pm$ standard deviation) spike volume was 3954.27 $\pm$ 863.50 mm^3 ($n = 477$) for early sampling, 5406.31 $\pm$ 1171.49 mm^3 ($n = 297$) for mid sampling, and 4961.13 $\pm$ 1105.81 mm^3 ($n = 360$) for late sampling (Fig. S2). The predominantly nadir view of the RGB cameras resulted in partial field point clouds, while the indoor scanner resulted in more detailed point clouds (Fig. S2). The dataset was split into training, validation, and test sets at the genotype level to prevent overfitting on genotype-specific features, ensuring that genotypes in the test set were not seen during training.

B. Evaluation of Spike Pairing

As there was no ground truth data available for the spike pairing, the method was evaluated quantitatively on artificial data. Specifically, a virtual scene was created containing 12 cameras, with the same calibration as the real cameras in the field. 800 of the scanned spikes were added to the scene, with distance, size and pose similar to how they occurred in the real data, resulting in 12 images containing around 400-500 spikes on each image (Fig. S4). The difference in number occurred, since some spikes were only visible on a subset of the cameras. The synthetic scenes were generated using Pyrender 0.1.45 [75]. Regarding the pose of the spikes, two different configurations were used: one where the spikes were mostly upright and rotated randomly with the angle being a normal distribution, and one where the angle was instead chosen uniformly. The first configuration was more similar to the scans occurring during earlier growing stages, where spikes were mostly upright, the later one was similar to later stages, where spikes tended to be more horizontal or random (Fig. S5).

Some spikes only appeared on a subset of the cameras. Those spikes were close to the border of the visible area. As they in fact often overlapped the border of the images, those could be seen as incorrect detections. We evaluated performance once with those spikes and once excluding them.

Accuracy was defined based on pairs of bounding boxes. A true positive occurred when two bounding boxes, belonging to the same spike, were correctly predicted as such. False negatives and false positives were defined analogously (Table S1). Averaged across 20 artificial sets of images, the performance dropped as fewer observations of a spike were present. Also when spikes at the border were included, recall dropped, while precision remained largely unaffected. This pattern was caused by many of the border spikes not being fully visible, hence those clusters broke into multiple smaller clusters, which resulted in many false negatives. By ignoring small clusters it was possible to filter out wrong or incomplete detections.

Table S1. Precision and recall for different cases of spike pairing. The best result in each category is highlighted.

Case	Views	Precision	Recall
Upright, Exclude border	6-9	0.95	0.87
	10-12	0.98	0.93
	12	**0.98**	**0.94**
Upright, Include all	6-9	0.96	0.80
	10-12	0.97	0.86
	12	**0.98**	**0.93**
Random, Exclude border	6-9	0.89	0.84
	10-12	0.94	0.91
	12	**0.95**	**0.93**
Random, Include all	6-9	0.91	0.72
	10-12	**0.95**	0.83
	12	**0.95**	**0.92**

C. Metric Calibration

Since volume is scale-dependent, accurate metric calibration was required for spike pairing, 3D reconstruction, and distance estimation. Camera poses were first estimated using openMVG [76] via Structure-from-Motion (SfM), which recovers geometry only up to an unknown global

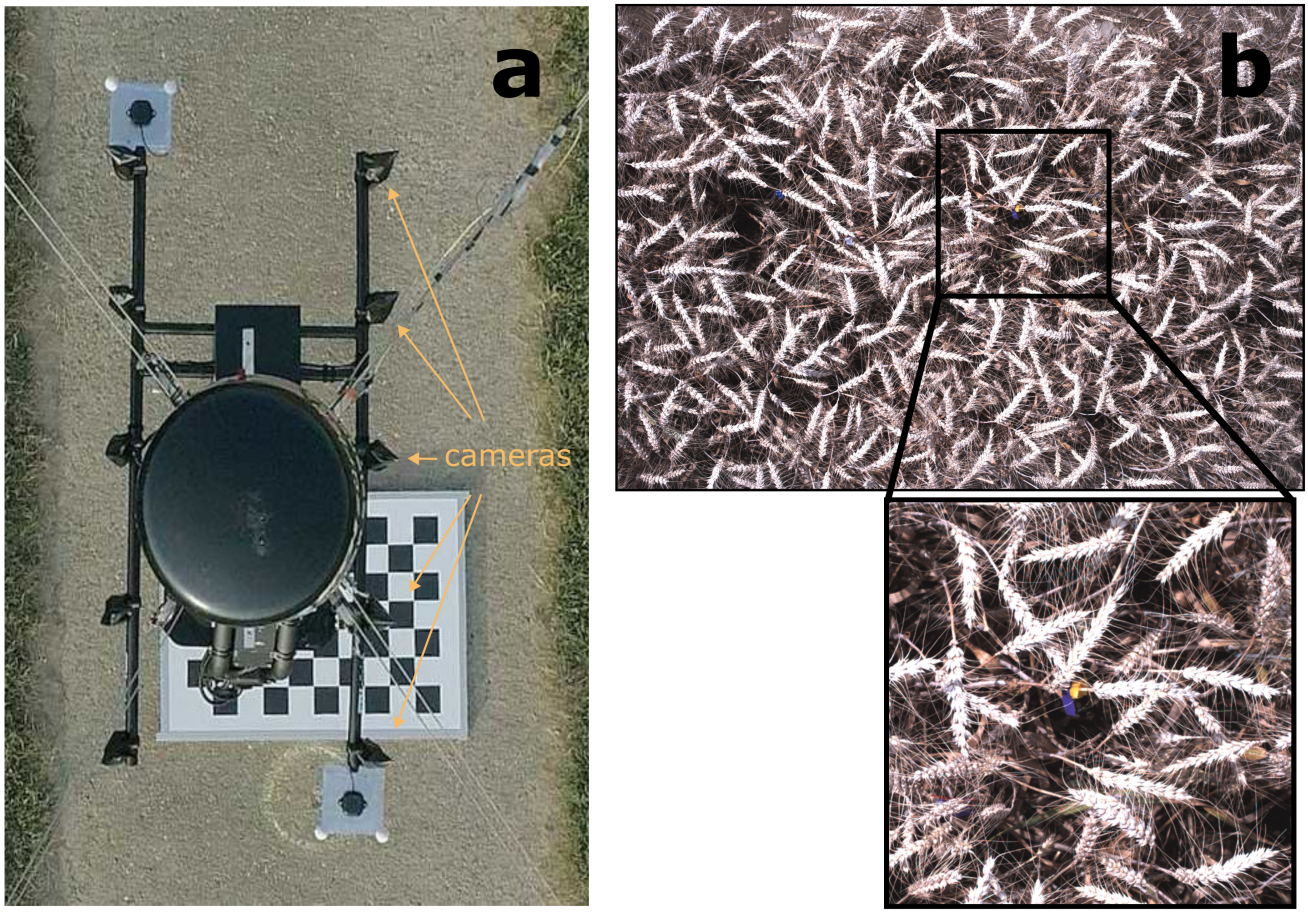

Figure S1. Field sensor equipped with 12 RGB cameras (a), and an example genotype plot with tagged spikes (b). Five cameras are positioned on each side (left and right, indicated by arrows), with two additional cameras located beneath the sensor.

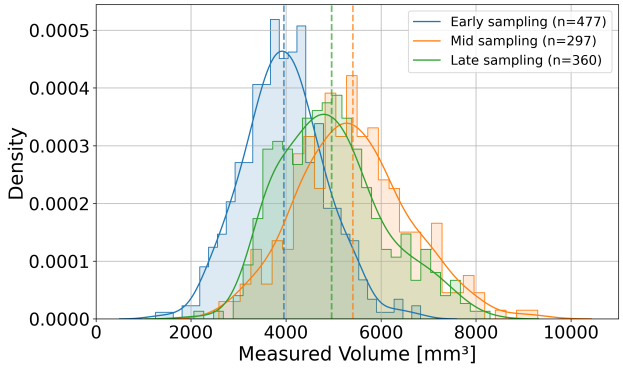

Figure S2. The volume distribution of the sampled spikes in 2023 and 2024, separated by the sampling growth stages, with the corresponding number of sampled spikes. The dashed line indicates the mean of each distribution.

scale. To receive metric scale, the SfM reconstructions were aligned with the metrically calibrated field phenotyping platform (FIP) system, whose camera extrinsics were determined using a checkerboard-based calibration, by matching pairwise inter-camera distances.

Let $c \in \{1, \ldots, M\}$ denote the camera index, where $M = 12$ cameras were used. SfM yields camera rotations R_c^{SfM} and camera centers $C_c^{\mathrm{SfM}} \in \mathbb{R}^3$, defined only up to scale, while the FIP calibration provides metrically scaled camera centers C_c^{FIP}. Pairwise inter-camera distances between cameras c and c', where $c' \neq c$ denotes a second camera index, were computed as $d_{cc'}^{\mathrm{FIP}} = \|C_c^{\mathrm{FIP}} - C_{c'}^{\mathrm{FIP}}\|_2$ and $d_{cc'}^{\mathrm{SfM}} = \|C_c^{\mathrm{SfM}} - C_{c'}^{\mathrm{SfM}}\|_2$. The global scale factor was estimated by averaging the ratios $d_{cc'}^{\mathrm{FIP}}/d_{cc'}^{\mathrm{SfM}}$ over all camera pairs (c, c'), i.e.,

$$s = \frac{1}{N} \sum_{c \neq c'} \frac{d_{cc'}^{\mathrm{FIP}}}{d_{cc'}^{\mathrm{SfM}}}, \tag{S1}$$

where $N = M(M-1)$ denotes the total number of ordered camera pairs. The SfM reconstruction was then metrically rescaled according to

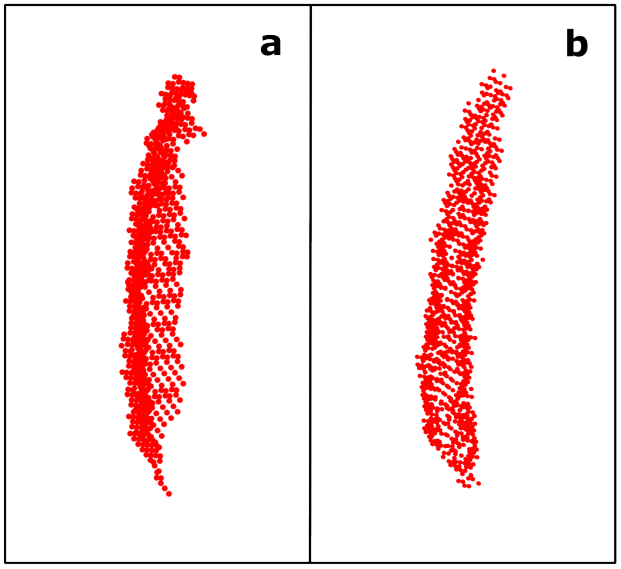

Figure S3. Example of a field point cloud (a), and an indoor scanned spike (b).

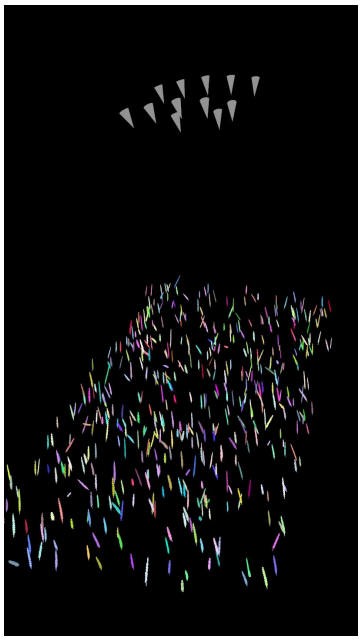

Figure S4. The setup of the artificial scene used for spike pairing evaluation. The cones represent the cameras.

$$C_c^{\text{metric}} = sC_c^{\text{SfM}}. \tag{S2}$$

The resulting multi-view clusters were then triangulated to estimate the 3D position and distances between camera and spikes. To estimate camera-to-spike distances, a 3D spike position was reconstructed for each multi-view cluster via repeated triangulation and its Euclidean distance to each camera was subsequently computed. Specifically, for a cluster observed in multiple images, R 3D points $X_r \in \mathbb{R}^3$, $r = 1, \ldots, k$ = 100 were obtained by again randomly sampling image points within the corresponding bounding boxes across a set of views and triangulated using the calibrated camera poses. Each triangulated point was weighted by its reprojection consistency across views. The spike center was then approximated as the weighted mean

Figure S5. Comparison of arbitrarily rotated (a), and upright spikes (b).

$$\bar{X} = \frac{1}{\sum_r w_r} \sum_{r=1}^{k} w_r X_r, \tag{S3}$$

where w_r denotes the fraction of views in which the triangulated point reprojects inside the corresponding bounding box. The distance to camera c was then computed as $d_c = \left\| \bar{X} - C_c^{\text{metric}} \right\|_2$, where C_c^{metric} denotes the metrically scaled camera center.

D. Loss Functions

D.1. Regulated Transformer

The regulated Transformer was trained with a regulated loss that combined image-level and spike-level supervision. For each spike s with n_s available views, we minimized

$$\begin{aligned} \mathcal{L}_{\text{RT},s} = \frac{1}{n_s} \sum_{j=1}^{n_s} \mathcal{L}_{\text{NLL}}(\mu_{j,s}, \sigma_{j,s}, v_s) \\ + 0.5\, \mathcal{L}_{\text{NLL}}(\mu_s, \sigma_s, v_s) \end{aligned} \tag{S4}$$

where

$$\mathcal{L}_{\text{NLL}}(\mu, \sigma, v) = \frac{(v-\mu)^2}{2\sigma^2} + \frac{1}{2} \log \sigma^2 \tag{S5}$$

denotes the Gaussian negative log-likelihood, $\mu_{j,s}, \sigma_{j,s}$ are the per-image predictions, and μ_s, σ_s are the spike-level predictions obtained from the Transformer volume token. The total loss was averaged over spikes in the batch.

$$\mathcal{L}_{\text{RT}} = \frac{1}{S} \sum_{s=1}^{S} \mathcal{L}_{\text{RT},s}. \tag{S6}$$

During validation and test time, only the global spike-level prediction μ_s was used as the final volume estimate.

D.2. Point Cloud Models

The point cloud models were optimized using a mean squared error (MSE) loss

$$\mathcal{L}_{\text{MSE}} = \frac{1}{S}\sum_{s=1}^{S}(\hat{v}_s - v_s)^2, \tag{S7}$$

where $\hat{v}_s$ and v_s denote predicted and ground-truth spike volumes. The loss was averaged over all spikes in the batch.

D.3. Distilled rigid-invariant Point Cloud Model

For the student rigid-invariant network trained on field-based point clouds, we minimized a combined loss

$$\mathcal{L}_{PC} = \mathcal{L}_{\text{MSE}} + \lambda\mathcal{L}_{\text{KD}}, \tag{S8}$$

where the feature distillation term is defined as

$$\mathcal{L}_{\text{KD}} = \frac{1}{S}\sum_{s=1}^{S}\left[\left\|\hat{z}_s^{\text{dir}} - z_s^{\text{dir}}\right\|_2^2 + \alpha\left(\|\hat{z}_s\|_2 - \|z_s\|_2\right)^2\right] \tag{S9}$$

with

$$z_s^{\text{dir}} = \frac{z_s}{\|z_s\|_2} \tag{S10}$$

and

$$\hat{z}_s^{\text{dir}} = \frac{\hat{z}_s}{\|\hat{z}_s\|_2}. \tag{S11}$$

Here, $\hat{z}_s$ and z_s denote the student and teacher latent features of the same spike s. The distillation loss was likewise averaged over spikes. We set $\lambda = 5$ and $\alpha = 0.2$.

D.4. Feature-distilled RT

For the distilled RT, we optimized a feature-based distillation that encouraged the RT to reproduce the latent representations of the multi-modal ensemble teacher. Training minimized the combined loss

$$\mathcal{L}_{\text{img}} = \mathcal{L}_{\text{RT}} + \beta\,\mathcal{L}_{\text{KD}}^{\text{feat}}. \tag{S12}$$

The feature distillation term was defined as

$$\mathcal{L}_{\text{KD}}^{\text{feat}} = \frac{1}{S}\sum_{s=1}^{S}\left[\left\|\hat{h}_s^{\text{dir}} - h_s^{\text{dir}}\right\|_2^2 + \gamma\left(\|\hat{h}_s\|_2 - \|h_s\|_2\right)^2\right] \tag{S13}$$

with

$$h_s^{\text{dir}} = \frac{h_s}{\|h_s\|_2} \tag{S14}$$

and

$$\hat{h}_s^{\text{dir}} = \frac{\hat{h}_s}{\|\hat{h}_s\|_2}. \tag{S15}$$

where $\hat{h}_s$ and h_s denote the projected student and teacher feature representations of spike s. The distillation loss was likewise averaged over spikes. We set $\beta = 0.2$ and $\gamma = 0.2$.

E. Runtime Performance

Runtime performance for the RT, the rigid-invariant PointNet for field-based point clouds, and the ensemble model was evaluated for the pre-processing steps, training and inference (Table S2). The pre-processing time for detection, pairing, and segmentation was identical across models. In contrast, both training and inference times differed, primarily due to backbone feature extraction in the image models and 3D reconstruction in the point-cloud-based approaches.

Table S2. Runtime comparison between the distilled regulated Transformer (RT), the rigid-invariant point cloud model (PC), and the ensemble.

Step	RT	PC	Ensemble
Image Pre-processing (seconds per plot)			
Detection	5.4	5.4	5.4
Pairing	7.0	7.0	7.0
Segmentation	0.46	0.46	0.46
Models (seconds per spike)			
Training	0.00184	0.16269	0.16244
Inference	0.00138	0.16053	0.16030

When estimating volume for many genotypes throughout an entire season, the distilled image model substantially reduced inference time compared to the point cloud models (Table S3).

Table S3. Runtime comparison between the distilled regulated Transformer (RT), and the point cloud models (PC), with an inference time of 1.4 ms for the RT and 160 ms for the PC models. Inference time was calculated for a genotype plot with 500 spikes, a field with 800 genotype plots, and 16 imaging time points per season, and reported in seconds (s), hours (h), and days (d).

	RT	PC Model
500 spikes (1 genotype plot)	0.7 s	80 s
800 plots (1 field)	0.16 h	17.78 h
16 time points (1 year)	2.49 h	11.85 d