Deep learning based 3d reconstruction for phenotyping of wheat seeds: a dataset, challenge, and baseline method

Vsevolod Cherepashkin
Forschungszentrum Jülich, IBG-2
Jülich, Germany

Erenus Yildiz
Forschungszentrum Jülich, IAS-8
Jülich, Germany

Andreas Fischbach
Forschungszentrum Jülich, IBG-2
Jülich, Germany

Leif Kobbelt
RWTH Aachen University, VCI
Aachen, Germany

Hanno Scharr
Forschungszentrum Jülich, IAS-8
Jülich, Germany

Abstract

We present a new data set for 3d wheat seed reconstruction, propose a challenge, and provide baseline methods. Individual plant seed properties influence early development of plants and are thus of interest in plant phenotyping experiments. Seed shape can be measured reliably from images using volume carving, as done in robotic setups such as phenoSeeder. However, about 36 images are needed to obtain a suitably accurate 3d model [33], where image acquisition takes $\approx 20\,s$. *For large-scale experiments with thousands of seeds higher throughput is required limiting image acquisition time. We present a deep-learning model that reconstructs an approximate 3d point cloud from fewer images, even only a single view. It has a significantly lower error than linear regression, which has been actively used so far in similar tasks. Using three images reduces imaging time by a factor of 10×, where relative errors of volume length, width, and height are all around* $2\,\%$. *Inference time from the neural network is negligibly short compared with imaging time which enables this method for real-time measurements and sorting.*

1. Introduction

Plant phenotyping investigates how expressed plant properties depend on their genotype and environmental conditions, as needed for breeding for novel traits in challenging climate and environmental conditions [11]. Different plant organs are usually of interest like roots [29, 48, 1], leaves [25, 2, 24, 12], flowers [50], fruits [26, 9, 10], or seeds [7] (see, e.g., [39, 36] for recent overviews on computer vision challenges and solutions in plant phenotyping).

Seed phenotyping allows quantifying seed properties and, e.g., relate them to germination properties or early plant development [6, 28, 35], or perform quality assessment and classification [49, 21]. In scientific or breeding experiments sorting seeds by properties, e.g., their size, can reduce or explain observed biological variation. Different seed phenotyping techniques are well-established. An acoustic volumeter [40, 41] has been developed for seeds. It directly measures seed volume excluding micro-pores, without shape reconstruction. When interested in shape, other methods are more suitable. Flatbed scanners have often been selected for seed 2d phenotyping [27, 47, 15]. While being technically simple and robust, they do not give full access to the 3d geometrical parameters of a seed, and different 3d seed phenotyping methods are available. Legume seeds were bulk phenotyped in 3d with a hand-held 3d laser scanner [16], being suitable for seeds in the several millimeters to centimeter range and reported to be imprecise for asymmetrical seeds. An automated seed phenotyping system called *pheno*Seeder [17] addresses also small seeds in the sub-millimeter range. It uses a single camera, a rotating robot arm, and volume carving (VC) [33] to reconstruct 3d voxel models of individual seeds [17]. They are used for volume estimation and, together with single seed mass measurements, allow for estimating mass density per seed.

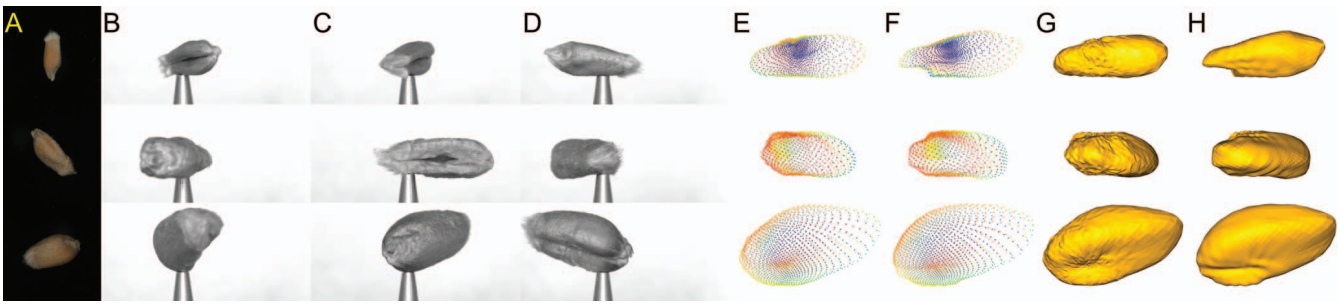

Figure 1: Example imaging data, ground truth VC point clouds and our method's predictions. Rows: Seeds with small, medium, and large volume. Columns: A: top view from '2d station'; B-D: three views out of 36 from '3d station'; E: point clouds of prediction; F: resampled VC point clouds used as ground truth (*gt*); G: meshes of prediction, H: meshes of *gt*.

VC is a shape-from-silhouette method [23, 32, 20, 4, 33] being robust under strongly varying seed surface properties resulting in, e.g., specular reflections, low contrast, etc. VC has also been applied for plant shoot 3d reconstruction [34, 13]. As this method does not use prior information on admissible shapes, reconstruction quality critically depends on a sufficiently high number of images of the object to be reconstructed. A quantitative analysis of spheres reveals that using 3 views results in approx. $+10\,\%$ systematic volume error, dropping to approx. $0.1\,\%$ for 35 or 36 images, depending on the camera configuration [33]. In *pheno*Seeder typically 36 images per seed are acquired at 2 images per second due to conceptual constraints on the robotic imaging system, consuming 18 s overall. This acquisition time limits throughput in large-scale experiments with thousands of seeds. Here, we aim at increasing throughput by reducing the number of images needed and thus acquisition time, while keeping reconstruction error reasonably small.

Neural networks have been proven to be practical for image-based 3d reconstruction in different representations and network types, e.g., voxel representations generated by deep convolutional neural networks (dCNN) [5] or transformers [44], triangular meshes by graph neural networks [45]. As deep learning-based methods can learn the space of admissible shapes, the same reconstruction quality as VC may be achieved using fewer images. In addition, once trained, reconstruction time needed by a dCNN model is considerably shorter than for VC at a sufficiently high resolution. It makes dCNN models to be a reasonable choice for large-scale plant seed phenotyping in 3d. As the seeds addressed here are simple, mostly convex objects, they lie in a relatively low-dimensional shape space and therefore should be reconstructable with an acceptably low error.

Deep CNNs have already been effectively applied in seed 2d-phenotyping for various tasks: classification [21, 22], instance segmentation [42], seed instance detection [46], prediction of the number of seeds [43]. However, to the best of our knowledge, deep learning-based methods have not yet been applied to 3d plant seed reconstruction. In particular, we use a white-box approach to estimate seed parameters like volume, width, height, and length. This means the 3d seed shape is reconstructed first and is thus available for human visual inspection if desired. The parameters of interest are derived in a subsequent step. This is in contrast to 'black-box' approaches, where parameters of interest are regressed directly from the input images. The validity of such black-box estimates is not easily verifiable, a well-known drawback of such deep learning solutions.

Our contribution

Dataset: We provide the train/val/test split of a new dataset with 2964 seeds as described in Section 2.2. It contains per seed: 1 color image, 36 gray-scale images from a turntable setup, 1 high-res 3d point cloud ($\approx$50k points) and 36 resampled point clouds (2k points each) of the seed's surface.

Challenge: We propose a challenge to derive a method with minimum relative seed volume error and/or minimal seed shape deviation using 1 or 3 input images (see Section 2.5). Evaluation is done on the test set with withheld ground truth.

New baseline method: We developed, investigated, and deployed a modified VGG11 [38] and ResNet-152 [14] for 3d reconstruction of plant seeds. It takes a few images as input and predicts a 3d point cloud of the seed's surface, and is functional even with only a single view per seed. Different image configurations have been tested and several geometrical metrics tracked. The model reaches an accuracy of shape and volume estimation being high enough for many plant phenotyping applications. It can be used in *pheno*Seeder or similar setups, to reduce measurement time by a factor of 10 or more compared to imaging requirements for VC, at the cost of statistical error increased to $\approx 2\,\%$.

2. Materials and methods

2.1. Raw Data

The raw dataset consists of 3357 wheat seeds, acquired on *pheno*Seeder (see Figure 1 for examples). In this sys-

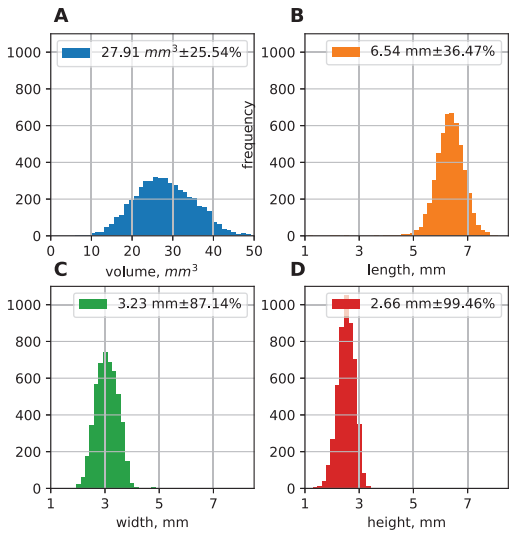

Figure 2: Distribution of the seed volumes (**A**), lengths (**B**), widths (**C**) and heights (**D**) across the dataset.

tem multiple seeds are presented on a glass plate, i.e., the so-called '2d imaging station' or short '2d station', where 8Mpix rgb-color 2d images are captured to locate the next seed to pick. They are cropped around the seed selected for pickup and offer a top-view on the seed (Fig. 1A). The robot picks the seed and moves it in the focal plane of a single static monochromatic camera of the so-called '3d station'. 36 gray-scale images (Fig. 1B) per seed are then captured while the seed is rotated in the horizontal plane by 10° steps in front of a featureless white background. These images are used for 3d reconstruction with VC. In [33] the robot's tool-tip was reconstructed together with the seed intentionally, and separated in 3d to prevent additional seed material removal. Point clouds derived by VC from all 36 images per seed were used to prepare ground truth (*gt*) data. The original voxel object was converted to a point cloud representing its surface only. The surface typically consists of ≈ 50000 points with a resolution of $\approx 0.05\,\text{mm}$ between neighboring points (Fig. 1D shows a subsampled version).

The 3d reconstruction result of VC was also used to estimate the ground truth volume V_{gt} of the seeds. In voxel representation, the number of voxels equals full volume. Distributions of seed volumes V, lengths L, widths W and height H, their respective means, and standard deviations in % are shown in Figure 2. Seed length L, width W and height H were estimated as main axes of an ellipsoid fitted to the predicted point cloud. The average single seed volume $\bar{V}$ over the whole dataset was $\bar{V} = 27.91\,\text{mm}^3$ with a relative standard deviation of $25.54\,\%$, and an overall spread from largest to smallest seed volume of about a factor of 4.5. Additionally, per seed mass measurements are available, measured with a high-end laboratory scale.

2.2. Prepared Datasets

Images: We scaled, shifted and cropped the raw data from the 3d station such that the position of the rotation axis and tool center point are static in all views. This makes dCNN training easier as it allows to use one fixed projection matrix as a good approximation to the real settings. We used image width $\times$ height being 373×200 pixels in our experiments.

A 'side view' was selected among the 36 images for every seed from the 3d station, being the image with the maximum visible foreground area derived from threshold segmentation. The 'tip view' of the same seed is the image captured at 90° after the side view. We used 2d images of the dataset in different ways as input data in our experiments:

- **all views:** one of 36 views are randomly drawn at each training iteration time as first view,
- **robot views:** the first view of the 36 views, irrespective of the orientation of the seed,
- **side views:** the 'side view' (see above) is used as first view,
- **tip views:** the 'tip view' is used as first view,
- **2d station views:** only the view from the 2d station is used.

In all but the last way (2d station views) multiple view configurations can be used.

Ground truth point clouds: We fixed the 3d origin of the ground truth VC point cloud to be at a constant position at a small distance above the tool center point lying reliably inside every seed. For wheat seeds we observe that they are in very good approximation star-shaped objects, where the origin lies inside the surface, and it is possible to connect every point on the surface with a straight line without crossing the surface elsewhere. This allows resampling of the seed surface across all the samples using spherical harmonics (SPH). To do so, we interpolated the ≈ 50000 discrete seed surface points derived with VC by least-squares fitting SPH using an open-source library [30, 31]. We found that a maximum SPH degree of $\ell = 20$ was sufficient to represent the *gt* surfaces in high enough detail (see Figure 3), which means $(20-1)^2 = 441$ SPH functions were used. This continuous interpolation allows sampling the seed surface in arbitrary directions from the seed origin. In our experiments, if not stated differently, we used a set D of 2000 new, fixed directions for the resampling. They were defined once and for all seeds to be the same. However, they need to be rotated according to the initial view of the 2d image used as input for the NN, such that the 3d coordinate system of the seed coincides with the 3d camera coordinate system of that view (or of the first view if multiple views are used as input to the NN, i.e., channel 0 of the input tensor). The directions

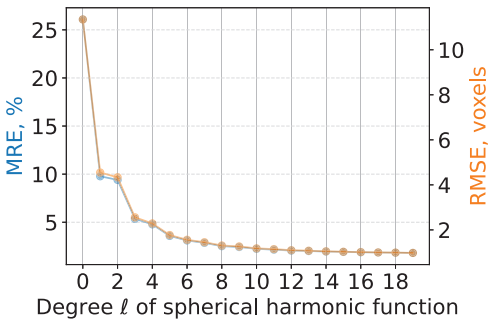

Figure 3: Fitting error vs maximum degree of spherical harmonic functions.

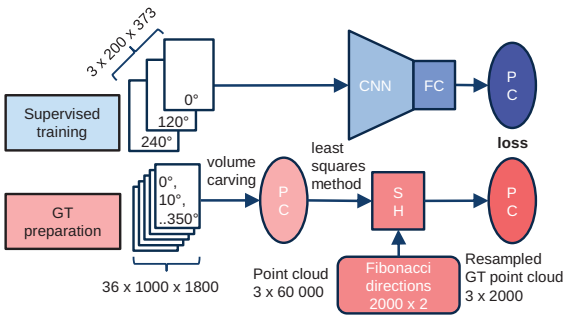

Figure 4: Setup: The dCNN receives several input images and returns a point cloud with known pairwise correspondences to the ground truth. Spherical harmonics are used to resample the VC ground truth point cloud with a fixed set of directions in 3d camera coordinates of the first input view.

were designed to be evenly distributed over a sphere using Fibonacci's golden ratio [19].

Outlier detection: In a preprocessing step outliers were filtered from the dataset, e.g., missing objects on the images (dropped seed) or impaired *gt* point cloud (incomplete or otherwise bad VC or SPH result). Local curvature thresholding and check on watertightness were used to filter these outliers. This results in $\approx 12\,\%$ of outliers and covers all observed error scenarios. After filtering 2964 of the initial 3357 seeds remained.

Train/val/test split: The dataset was split into train/val/test: $70/15/15\,\%$. In the beginning, the dataset was randomly shuffled and the first $15\,\%$ of the data was separated as a test set ($K = 444$ samples after filtering). Then the rest was randomly shuffled four times, for each time train/val sets were separated. Experiment results shown below in Section 3 are for the test set.

Data provided: We provide the train/val part of the filtered dataset with 2520 seeds. For the test split only 3 images per seed are provided ('robot view'), ground truth is held back. The dataset contains 2d station and 3d station images, stabilized and resampled crops, 1 full resolution and 36 resampled point clouds, i.e., 1 per 3d station view.

2.3. Losses and performance metrics for shapes

The neural network (NN) receives N 2d images as input and regresses a 3d point cloud. Every point of the generated point cloud is represented as a vector with fixed preset direction but variable length, see **Figure 4**. Consequently, the NN just needs to regress the lengths of the vectors. This is a simplification wrt. general point cloud based surfaces that is possible due to the approximate star-shapedness of wheat seeds. This allows to use simple L_p-norm like loss functions

$$L_p = \frac{1}{M}\left(\sum_{i=1}^{M} |\mathbf{v_i} - \mathbf{v_i^{gt}}|^{\mathbf{p}}\right)^{1/p} \tag{1}$$

where $M = 2000$ is the number of points in the point cloud, $\mathbf{v_i}$ predicted and $\mathbf{v_i^{gt}}$ ground truth (*gt*) vectors. We use the L_1 loss function for training, i.e., mean absolute difference of $\mathbf{v_i}$ and $\mathbf{v_i^{gt}}$. In experiments we also report reconstruction accuracy in terms of mean Euclidean distance L_2 and maximum per point difference (MPD, i.e., $M \cdot L_\infty$).

When surface point density varies significantly in different regions of the seed surface, point-based methods may lead to over- or under-representation of seed regions in the loss. Therefore, we also investigate a loss L_T based on surface triangulations T^{gt} and T of the ground truth and predicted point clouds, respectively. With L_T we aim to capture the difference between volumes $V_T(T_j^{gt})$ and $V_T(T_j)$ of tetrahedra spanned by the origin and corresponding triangles T_j^{gt} and T_j. In order to handle all special and degenerate cases robustly we make two modifications to the simplistic and insufficient $\tilde{L}_T = \sum_j |V_T(T_j^{gt}) - V_T(T_j)|$

$$L_T = \sum_j V_T(T_j^{\uparrow}) - c_j V_T(T_j^{\downarrow}) \tag{2}$$

where j enumerates all triangles, $T_j^{\uparrow}$ is the outer and $T_j^{\downarrow}$ the inner triangle derived like this: We use that corresponding points $\mathbf{v_i}$ and $\mathbf{v_i^{gt}}$ lie on the same ray and define $\mathbf{v_i^{\uparrow}} = \mathbf{v_i}$ if $|\mathbf{v_i}| > |\mathbf{v_i^{gt}}|$, else $\mathbf{v_i^{\uparrow}} = \mathbf{v_i^{gt}}$. Similarly, $\mathbf{v_i^{\downarrow}} = \mathbf{v_i}$ if $|\mathbf{v_i}| \leq |\mathbf{v_i^{gt}}|$, else $\mathbf{v_i^{\downarrow}} = \mathbf{v_i^{gt}}$. Let $\mathcal{T}_j$ be such that $i \in \mathcal{T}_j$ consistently enumerate the corner points of T_j^{gt} and T_j, respectively. $T_j^{\uparrow}$ is then given by $\mathbf{v_{i\in\mathcal{T}_j}^{\uparrow}}$ and $T_j^{\downarrow}$ by $\mathbf{v_{i\in\mathcal{T}_j}^{\downarrow}}$. Further, $c_j = 1$ if vectors $\mathbf{v_i^{gt}}$ and $\mathbf{v_i}$ are parallel for all $i \in \mathcal{T}_j$, else $c_j = -1$.

Chamfer distance [45, 8] and 3d-Intersection over Union (3d-IoU) [5] are frequently used as losses in 3d reconstruction tasks. Due to the resampling of the seed's surfaces, surface point pairs needed for Chamfer distance are known in advance, and it boils down to our simple L_2 loss.

2.4. Seed parameters and model calibration

Seed properties like length L, width W, height H, and volume V are the parameters of interest in biological applications. Therefore we investigate their mean absolute percentage errors, i.e., E_L, E_W, E_H, and E_V, respectively

$$E_y = \frac{1}{K} \sum_{i=1}^{K} \frac{|y_{\text{gt},i} - y_{\text{pred},i}|}{y_{\text{gt},i}} \tag{3}$$

with $y \in \{V, L, W, H\}$ and K being the number of seeds in the test split.

Seed volume was calculated from triangular meshes as a sum over all tetrahedra with the common fixed origin. As the set D of directions used for resampling of each seed surface is fixed, the triangulation of all seed surfaces is also fixed, i.e., meshes were derived using deformation of the unit sphere mesh with the same set D of directions.

Model calibration: Being a white-box approach to estimate seed parameters like volume, width, height and length, the proposed method is not necessarily bias free. In fact, using L_1 loss (1), we observe, e.g., a systematic underestimation of volume. We therefore calibrate the volume V_{pred} calculated from the shape prediction towards ground truth volume V_{gt}. This means, we fit a linear function $f(V_{\text{pred}})$ to the scatter plot of V_{gt} versus V_{pred} and convert predictions as $V = f(V_{\text{pred}})$. We do this for all derived models and also for L, W, and H.

For black-box models trained with a loss minimizing, e.g., volume directly, i.e., $V_{\text{pred}} - V_{gt}$ in some norm, such calibration is not really needed (cmp. Figure 9).

2.5. Challenge

The proposed challenge is to minimize E_V (3) and/or L_T (2) given either a single or three 3d station images. Training can be but does not have to be fully supervised using the *gt* point clouds. Using additional data is allowed, with the exception of the (unavailable) test split data. Automated evaluation on the blind test set is available online[1].

2.6. Simple baseline methods

Mean volume: The average single seed volume $\bar{V}$ over the whole dataset was $\bar{V} = 27.91\,\text{mm}^3$ with a relative standard deviation of $25.54\,\%$, as shown in **Figure 2A**. Consequently, a trivial seed volume estimator delivering $\bar{V}$ independent from the input has a mean absolute percentage error of $E_{\bar{V}} = 21.2\,\%$, being the lowest quality baseline to beat.

Volume from projected area: In [17] a well-established method for seed volume estimation from 2d images has been used as baseline, that VC clearly surpasses. They used 2d imaging station images to regress a volume value V_{2D} from the projected seed area A, i.e., the number of foreground pixels in the seed mask segmented from the background by simple thresholding. Volumes were predicted from the projected area according to

$$V_{\text{2D}} = c \cdot A^{3/2} \tag{4}$$

where the 'shape parameter' c was fit to the data and depends on seed species.

Direct regression of volume: Black-box approaches can be used as alternative to our white-box method with the same input view variants. We use ResNet-152 [14] as base model, but with the last layer appended with nn.Linear(5, 1) without activation function. We train it with standard L_2 loss on the ground truth volume V_{gt}, learning rate $3 \cdot 10^{-5}$.

2.7. Model architecture and training procedure

We use a simple VGG11 [38] with batch normalization as well as a ResNet-152 [14] as backbone models. The last fully connected layer was modified to have M output neurons, to fit the shape *gt* vector, as shown in **Figure 6** for the VGG11. Preliminary tests (not shown) minimizing L_1 metrics at fixed set of hyperparameters revealed that this small model has sufficient descriptive power for the task. However, better models may exist.

Multiple view handling: In the case of multiple input images per seed, views were aggregated in the color channel of the input tensor. The first of the images was assigned to a certain view. When allowing multiple input images per seed, the images were always given in a fixed configuration. This means, the images were taken at fixed angles $\phi(N, i)$ relative to the first image, where N is the number of images used per seed and i stands for the i^{th} image. The angles are given by

$$\phi(N, i) = 10 \cdot (\text{round}\ (18\ i\ (1 + \frac{1}{N})) \mod 36) \tag{5}$$

for $N \in \{1, ..., 36\}$, and for $i \in \{0, 1, .., N\}$, where round$(\cdot)$ returns the nearest integer. This means that for odd N, angles were distributed equidistantly. For even numbers semicircle was divided into equal sectors, and every second angle was flipped over $180°$, see Figure 5. The idea behind the 'flipped over' configuration is that for even N views are pairwise aligned on opposite sides of the seed and thus add only limited geometrical information. 'Flip over' configurations also cover the whole seed surface, but offer more diverse information.

Hyperparameters, software, and timing: Batch size 25 was used for all training runs and learning rate $3 \cdot 10^{-5}$. Architectures with various numbers of input views, i.e. $N \in [1, 3, 4, 5, 6, 9, 12]$, were tested.

The training was done on a server with four Nvidia A100 GPUs[2]. Data-parallel training with Horovod [37] for PyTorch[3] was used to distribute data over GPUs.

[1] https://helmholtz-data-challenges.de

[2] https://www.nvidia.com/en-us/data-center/a100

[3] https://pytorch.org/ version 1.12.1, py3.9_cuda11.6_cudnn8_0

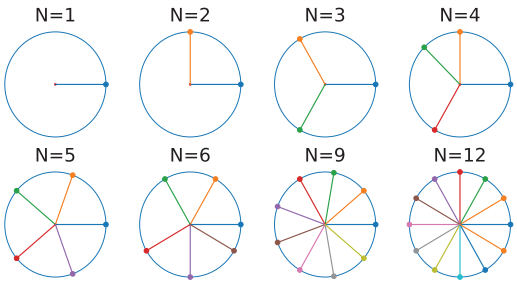

Figure 5: Selection of views for 3d reconstruction. Object is in the center in virtual setup and cameras are on the circle.

Figure 6: Architecture of VGG11 [38] with the modified first and last layer. N input color channels for multiple view configurations. The last layer outputs $M = 2000$ values, corresponding to the number of points in point cloud.

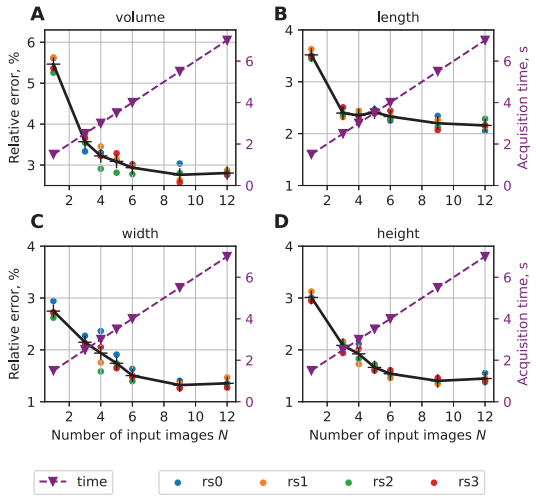

Figure 7: Dependency of the relative error E on the number of input images N (solid line) and comparison with acquisition time (dashed line). "rs" means different training runs with different random seeds.

Inference of a point cloud from the trained model was done on a single Nvidia A100 GPU. In batch-wise calculation including data input and output, a single seed point cloud was derived in 49 ± 1 ms per seed on average.

3. Results

From a biological point of view, seed volume is the geometric parameter of highest interest in typical plant experiments, as it is the best proxy for seed mass; and seed mass correlates to the amount of nutrients initially available to a seedling. Length, width, and height are less relevant but can also be used for plant phenotyping. In our experiments below, we therefore investigate model performance focusing on relative volume error and report other measures for completeness.

3.1. Selecting suitable view configurations

In order to find the best performing view configuration, we use the model and training procedure described in Section 2.7 and loss L_1. We trained all the different view configurations also described there, using the 'all views' dataset configuration (see Section 2.2). We repeated the training four times with different random seeds to evaluate training variation.

Relative errors E_V, E_L, E_W and E_H (see (3)) versus the number N of input images are shown in Figure 7. We observe in Figure 7A, that the relative volume error E_V ranges between $\lessapprox 5.5\,\%$ when using a single view and $\approx 2.8\,\%$ when using more than 9 views. For 3 to 6 views the error is around 3 % to 3.5 %. Variation between training runs with different random seeds is between ± 0.25 and ± 0.1 percent points. Similar but less pronounced behavior can be observed for length, width, and height in Figures 7B-D.

In addition to the errors, the figures also show the time needed to capture the images. Non-surprisingly, it increases linearly with the number of images captured, where some overhead is needed for moving the seed to the 3d imaging station. Considering our goal of speeding up imaging for trading some accuracy wrt. VC on 36 images, the best options seem to be: (1) using a single view, as it is the fastest option and still offers low enough error to be acceptable in high-throughput experiments with good number statistics; (2) using 3 views, as using 4 to 6 views takes more time but does not increase accuracy considerably. In scenarios, where the increased accuracy by using more views would be relevant, using full VC may be advisable.

In order to investigate the per point accuracy of the predicted point clouds, we calculate L_1 and L_2 metrics in mm, as well as mean maximum per point distance (MPD) in mm, and L_T in mm^3 for models trained on L_1 (see Figure 8). We see that the error values generally go down with increasing views N, with considerable variance at $N = 4$ and $N = 5$. Mean L_1 values for 3 views are $\approx 40\,\mu$m, i.e., the average distance of a predicted grid point to its ground truth counterpart. The MPD for the same configuration shows that the mean maximum outlying grid point is at $\approx 500\,\mu$m. Consequently the main error comes from relatively few far off points while the others are very close to the ground truth. We believe that this is due to the high point density close to the tool tip / origin, and that the L_T loss mitigates this drawback.

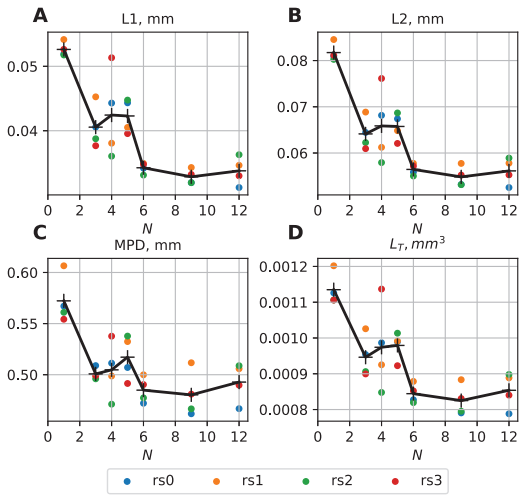

Figure 8: Metrics values L_1, L_2, MPD, and L_T vs. number N of input images. Black line: mean among different training runs; rs0 to rs3 indicate different random seeds.

3.2. Volume estimation comparison

We compare volume estimation results on the 'robot view' of the different variants of our method with usually employed baselines described in Section 2.6, i.e.:

- volume from projected area, using the '2d station view' (method M1) or the 'side view' (M2, see Section 2.2)
- direct 'black-box' regression of volume from 1 view (M3) or 3 views (M4) using ResNet-152,
- our 'white-box' method with VGG11, 1 view (M5) or 3 views (M6) trained with L_1 loss,
- our 'white-box' method with ResNet-152, 1 view (M7) or 3 views (M8) trained with L_1 loss from (1).
- our 'white-box' method with ResNet-152, 1 view (M9) or 3 views (M10) trained with L_T loss from (2).

The overall volume variation in our set in terms of relative error (3) is $E_{\bar{V}} = 21.2\ \%$, giving a reference on what to consider as a 'small' or 'large' error.

Figure 9 shows scatter plots of predicted volume before bias correction versus ground truth volume for the different methods. From the red dashed fit line we see that bias correction is less needed for methods M1 to M4 deriving volume directly, but is beneficial for our two-step white-box methods M5 to M10. Without this calibration M5 to M10 underestimate volume. As expected, methods with 3 input images perform more reliably than their single view counterparts. L_T based methods work better than L_1. ResNet-152 outperforms VGG11 for 3 views, while there is no benefit for 1 view. Single view direct regression is almost as good as the best single view white-box method M9, however with

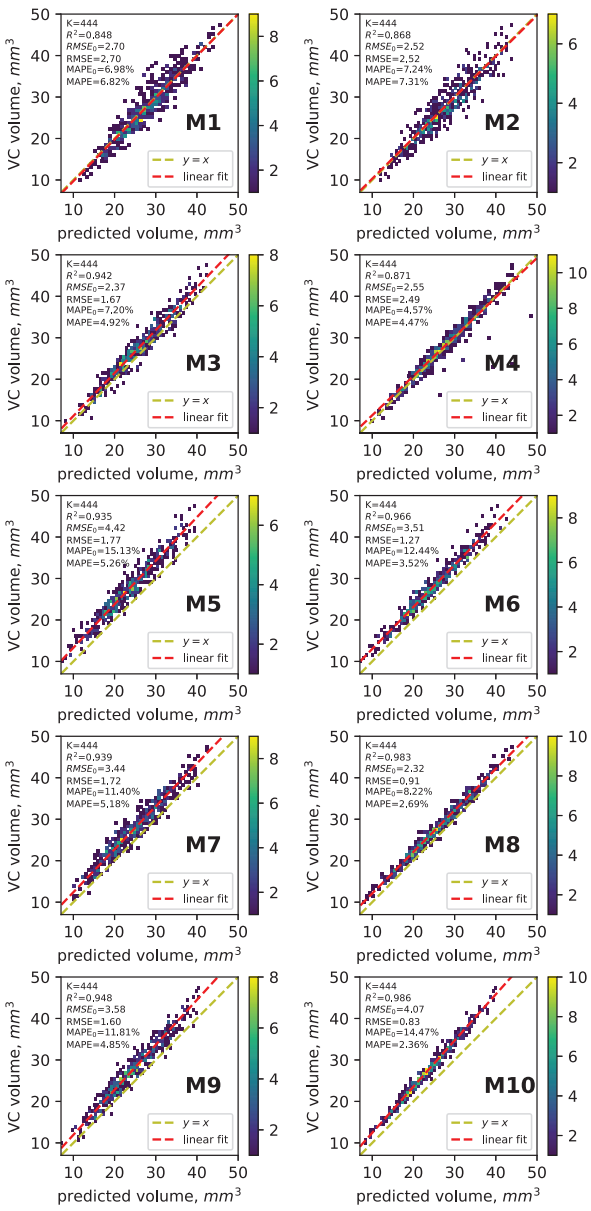

Figure 9: Volume from volume carving (V_{gt}) compared with predicted volume of methods M1 to M10 (see Section 3.2). MAPE_0 and MAPE - relative error, and RMSE_0 and RMSE - root mean squared error before and after correction, respectively. 2d histograms with K points, color bar shows number of samples per bin.

3 views all white-box methods beat direct regression. In Table 1 we see that quantitatively M10 performs clearly best with $E_V = 2.36\,\%$ only.

3.3. Influence of seed pose

Using our white-box approach with 1 and 3 views and L_1 loss, we investigate if starting imaging with a predefined

Model	N	E_V	E_L	E_W	E_H
M1, 2d station	1	6.82	4.66	4.84	5.1
M2, side view	1	7.31	–	–	–
M3, dir. regr., ResNet	1	4.92	–	–	–
M4, dir. regr., ResNet	3	4.47	–	–	–
M5, ours L_1,VGG11	1	5.26	3.49	2.61	2.96
M6, ours L_1,VGG11	3	3.52	2.36	**2.09**	2.1
M7, ours L_1, ResNet	1	5.18	3.76	2.83	2.99
M8, ours L_1, ResNet	3	2.69	2.59	2.38	**2.04**
M9, ours L_T, ResNet	1	4.85	3.41	3.18	3.07
M10, ours L_T, ResNet	3	**2.36**	**1.94**	2.47	2.14

Table 1: Relative errors [%] of predicted volume, length, width, and height using methods M1 to M10 (see Sec. 3.2).

		Training		
N	Test	**av**	**sv**	**tv**
1	rv	5.47±0.16	17.17±0.49	16.53±1.35
	sv	5.35±0.06	9.09±1.42	21.07±0.75
	tv	5.51±0.23	12.25±0.53	8.61±0.37
3	rv	3.57±0.15	15.98±1.73	10.10±0.43
	sv	3.46±0.16	8.34±1.63	13.26±0.57
	tv	3.41±0.34	15.51±1.10	7.57±1.32

Table 2: Relative volume error [%] for different train and test datasets. **av** – all, **sv** – side, **tv** – tip, **rv** – robot views (see Section 2.2). Columns indicate training, rows test scenarios. N is number of input views to the model.

seed pose would be beneficial. Such pose can in principle be derived from the 2d station view before picking up the seed. In addition to the models trained on all views, we trained models on side views only, and on tip views only. We trained for the same number of iterations as with 'all views'

We expected, that training and testing on side views only would yield best performance. Our intuition was that side views show most seed area and, therefore, most information for volume estimation. Further, keeping the input as constant as possible should allow the model to adapt best.

Table 2 reveals, that this intuition is wrong in several aspects. Training on 'all views' (first column) clearly outperforms training on 'side views' or 'tip views', independent of the test scenario (rows). Thus, keeping the seed pose fixed for training does not help. Further, we tested the model trained on all views for single views (N=1) for prediction with different seed poses, i.e., 'robot views' (see Section 2.2), side and tip views. Against our intuition, we do not see a significant error difference when using tip views vs. side views. While surprising at first glance, this result makes sense when considering seed property statistics from Figure 2. A side view shows length and height of a seed, but not its width, while a tip view shows width and height, but not its length. However, relative width variation of the investigated wheat seed species is higher (87 %) than length variation (36 %). Height is always visible in all views. Consequently tip views contain even more information on volume than side views and a significant error increase would not be plausible.

When using more than one view, i.e., $N = 3$ errors drop significantly by approx. 2% for the 'all views' training. For 'side view' and 'tip view' errors remain higher than when using a single view, i.e. $N = 1$, but training on 'all views'.

4. Discussion and Conclusion

The proposed deep learning based method for 3d reconstruction and phenotyping of individual plant seeds from single and multiple images can substitute volume carving, currently used in robot setups like *pheno*Seeder [17]. It allows to save acquisition time, as it can successfully operate with fewer images per seed than VC (e.g., 3 instead of 36). Especially in high-throughput experiments with thousands or tens of thousands of seeds, processing time per seed needs to be kept as low as possible. Trading acquisition time for some accuracy of seed parameter estimation and thus expressiveness is therefore acceptable. Three input views per seed are found to be a recommended trade-off between acquisition time (10× reduction) and reached accuracy of ≈ 2.36 % relative error. It can be used in turn-table setups with three poses as in *pheno*Seeder or in affordable setups with three cameras and simpler robotics and the same acquisition time as needed for a single image.

Even with a single view per seed the model infers a valid point cloud with lower error than readily available linear regression from projected area. The presented new baseline method therefore has potential for 3d reconstruction and approximate estimation of parameters without turn-table setup directly from a single 2d view, e.g., when seeds lie on a surface like the 2d station views. However, here, our simple loss requires the ground truth point cloud to be aligned with the pose in the first image. This is not given for the 2d station views. A more complicated loss function than the one used here, e.g., Chamfer loss, may thus be needed to train a well-performing model.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Authors' contributions: VC, LK, HS conceived the study, AF captured the data, VC and EY ran the experiments. All authors wrote and proofed the paper.

Acknowledgments: The authors gratefully acknowledge computing time on the supercomputer JURECA [18] at Forschungszentrum Jülich under grant delia-mp.

Data Availability Statement: The datasets presented here can be found in the EUDAT b2share repository [3].

References

[1] T. Bontpart, C. Concha, M. V. Giuffrida, I. Robertson, K. Admkie, T. Degefu, N. Girma, K. Tesfaye, T. Haileselassie, A. Fikre, M. Fetene, S. A. Tsaftaris, and P. Doerner. Affordable and robust phenotyping framework to analyse root system architecture of soil-grown plants. *The Plant Journal*, 103(6):2330–2343, 2020. 1

[2] A. Chaudhury, P. Hanappe, R. Azaïs, C. Godin, and D. Colliaux. Transferring pointnet++ segmentation from virtual to real plants. In *CVPPA-ICCV*, 2021. 1

[3] V. Cherepashkin, E. Yildiz, A. Fischbach, L. Kobbelt, and H. Scharr. Deep learning based 3d reconstruction for phenotyping of wheat seeds: dataset, 2023. EUDAT b2share repository, doi: 10.34730/b716920e6c4342879ddb46fbcfbafe17, https://b2share.fz-juelich.de. 8

[4] K. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages I–77–I–84, Madison, WI, USA, 2003. IEEE Comput. Soc. 2

[5] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction, Apr. 2016. 2, 4

[6] J. Colmer, C. M. O'Neill, R. Wells, A. Bostrom, D. Reynolds, D. Websdale, G. Shiralagi, W. Lu, Q. Lou, T. Le Cornu, J. Ball, J. Renema, G. Flores Andaluz, R. Benjamins, S. Penfield, and J. Zhou. SeedGerm: A cost-effective phenotyping platform for automated seed imaging and machine-learning based phenotypic analysis of crop seed germination. *New Phytologist*, 228(2):778–793, Oct. 2020. 1

[7] E. David, M. Serouart, D. Smith, S. Madec, K. Velumani, S. Liu, X. Wang, F. P. Espinosa, S. Shafiee, I. S. Tahir, et al. Global wheat head dataset 2021: more diversity to improve the benchmarking of wheat head localization methods. *arXiv preprint arXiv:2105.07660*, 2021. 1

[8] H. Fan, H. Su, and L. Guibas. A Point Set Generation Network for 3D Object Reconstruction from a Single Image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2463–2471, Honolulu, HI, July 2017. IEEE. 4

[9] Z. Fei, A. G. Olenskyj, B. N. Bailey, and M. Earles. Enlisting 3d crop models and gans for more data efficient and generalizable fruit detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1269–1277, 2021. 1

[10] M. Feldmann, A. Tabb, and S. Knapp. Cost-effective, high-throughput 3d reconstruction method for fruit phenotyping. *Computer Vision Problems in Plant Phenotyping (CVPPP)*, 1, 2019. 1

[11] F. Fiorani and U. Schurr. Future Scenarios for Plant Phenotyping. *Annual Review of Plant Biology*, 64(1):267–291, Apr. 2013. 1

[12] M. Gaillard, C. Miao, J. Schnable, and B. Benes. Sorghum segmentation by skeleton extraction. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, pages 296–311. Springer, 2021. 1

[13] M. Gaillard, C. Miao, J. C. Schnable, and B. Benes. Voxel carving-based 3D reconstruction of sorghum identifies genetic determinants of light interception efficiency. *Plant Direct*, 4(10), Oct. 2020. 2

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition, Dec. 2015. 2, 5

[15] R. P. Herridge, R. C. Day, S. Baldwin, and R. C. Macknight. Rapid analysis of seed size in Arabidopsis for mutant and QTL discovery. *Plant Methods*, 7(1):3, Dec. 2011. 1

[16] X. Huang, S. Zheng, and N. Zhu. High-Throughput Legume Seed Phenotyping Using a Handheld 3D Laser Scanner. *Remote Sensing*, 14(2):431, Jan. 2022. 1

[17] S. Jahnke, J. Roussel, T. Hombach, J. Kochs, A. Fischbach, G. Huber, and H. Scharr. *Pheno* Seeder - A Robot System for Automated Handling and Phenotyping of Individual Seeds. *Plant Physiology*, 172(3):1358–1370, Nov. 2016. 1, 5, 8

[18] Jülich Supercomputing Centre. JURECA: Data Centric and Booster Modules implementing the Modular Supercomputing Architecture at Jülich Supercomputing Centre. *Journal of large-scale research facilities*, 7(A182), 2021. 8

[19] B. Keinert, M. Innmann, M. Sänger, and M. Stamminger. Spherical fibonacci mapping. *ACM Transactions on Graphics*, 34(6):1–7, Nov. 2015. 4

[20] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, Feb./1994. 2

[21] A. Loddo, M. Loddo, and C. Di Ruberto. A novel deep learning based approach for seed image classification and retrieval. *Computers and Electronics in Agriculture*, 187:106269, Aug. 2021. 1, 2

[22] T. Luo, J. Zhao, Y. Gu, S. Zhang, X. Qiao, W. Tian, and Y. Han. Classification of weed seeds based on visual images and deep learning. *Information Processing in Agriculture*, 10(1):40–51, Mar. 2023. 2

[23] W. N. Martin and J. K. Aggarwal. Volumetric Descriptions of Objects from Multiple Views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):150–158, Mar. 1983. 2

[24] T. Masuda. Leaf area estimation by semantic segmentation of point cloud of tomato plants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1381–1389, 2021. 1

[25] M. Minervini, A. Fischbach, H. Scharr, and S. A. Tsaftaris. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern recognition letters*, 81:80–89, 2016. 1

[26] M. Miranda, L. Zabawa, A. Kicherer, L. Strothmann, U. Rascher, and R. Roscher. Detection of anomalous grapevine berries using variational autoencoders. *Frontiers in Plant Science*, 13, 2022. 1

[27] F. B. Musaev, N. S. Priyatkin, M. I. Ivanova, A. F. Bukharov, and A. I. Kashleva. Analysis of morphometric and optical parameters of seeds of the subgenus cepa (*Allium* L., Alliaceae) by digital scanning. *Siberian Herald of Agricultural Science*, 52(2):22–31, May 2022. 1

[28] D. Paczesniak, M. Pellino, R. Goertzen, D. Guenter, S. Jahnke, A. Fischbach, J. T. Lovell, and T. F. Sharbel. Seed size, endosperm and germination variation in sexual and apomictic Boechera. *Frontiers in Plant Science*, 13:991531, Nov. 2022. 1

[29] D. Pflugfelder, J. Kochs, R. Koller, S. Jahnke, C. Mohl, S. Pariyar, H. Fassbender, K. A. Nagel, M. Watt, and D. van Dusschoten. The root system architecture of wheat establishing in soil is associated with varying elongation rates of seminal roots: quantification using 4d magnetic resonance imaging. *Journal of experimental botany*, 73(7):2050–2060, 2022. 1

[30] A. Politis. Real/complex spherical harmonic transform, gaunt coefficients and rotations, 2023. Retrieved June 30, 2023. 3

[31] A. Politis et al. *Microphone array processing for parametric spatial audio techniques*. Aalto University, 2016. 3

[32] M. Potmesil. Generating octree models of 3D objects from their silhouettes in a sequence of images. *Computer Vision, Graphics, and Image Processing*, 40(1):1–29, Oct. 1987. 2

[33] J. Roussel, F. Geiger, A. Fischbach, S. Jahnke, and H. Scharr. 3D Surface Reconstruction of Plant Seeds by Volume Carving: Performance and Accuracies. *Frontiers in Plant Science*, 7, June 2016. 1, 2, 3

[34] H. Scharr, C. Briese, P. Embgenbroich, A. Fischbach, F. Fiorani, and M. Müller-Linow. Fast High Resolution Volume Carving for 3D Plant Shoot Reconstruction. *Frontiers in Plant Science*, 8:1680, Sept. 2017. 2

[35] H. Scharr, B. Bruns, A. Fischbach, J. Roussel, L. Scholtes, and J. vom Stein. Germination Detection of Seedlings in Soil: A System, Dataset and Challenge. In A. Bartoli and A. Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, volume 12540, pages 360–374. Springer International Publishing, Cham, 2020. 1

[36] H. Scharr and S. A. Tsaftaris. Meeting computer vision and machine learning challenges in crop phenotyping. In *Advances in plant phenotyping for more sustainable crop production*. Burleight Dodds, 2022. 1

[37] A. Sergeev and M. Del Balso. Horovod: Fast and easy distributed deep learning in TensorFlow, Feb. 2018. 5

[38] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, Apr. 2015. 2, 5, 6

[39] I. Stavness, V. Giuffrida, and H. Scharr. Editorial: Computer vision in plant phenotyping and agriculture. *Frontiers in Artificial Intelligence*, 6, 2023. 1

[40] Sydoruk, Kochs, van Dusschoten, Huber, and Jahnke. Precise Volumetric Measurements of Any Shaped Objects with a Novel Acoustic Volumeter. *Sensors*, 20(3):760, Jan. 2020. 1

[41] V. Sydoruk, J. Kochs, D. van Dusschoten, and S. Jahnke. Device and method for determining the volume and porosity of objects and bulk materials, 2019. US patent US20220011272A1, pending. 1

[42] Y. Toda, F. Okura, J. Ito, S. Okada, T. Kinoshita, H. Tsuji, and D. Saisho. Training instance segmentation neural network with synthetic datasets for crop seed phenotyping. *Communications Biology*, 3(1):173, Apr. 2020. 2

[43] L. Uzal, G. Grinblat, R. Namías, M. Larese, J. Bianchi, E. Morandi, and P. Granitto. Seed-per-pod estimation

for plant breeding using deep learning. *Computers and Electronics in Agriculture*, 150:196–204, July 2018. 2

[44] D. Wang, X. Cui, X. Chen, Z. Zou, T. Shi, S. Salcudean, Z. J. Wang, and R. Ward. Multi-view 3D Reconstruction with Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5702–5711, Montreal, QC, Canada, Oct. 2021. IEEE. 2

[45] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, volume 11215, pages 55–71. Springer International Publishing, Cham, 2018. 2, 4

[46] P. Wang, F. Meng, P. Donaldson, S. Horan, N. L. Panchy, E. Vischulis, E. Winship, J. K. Conner, P. J. Krysan, S.-H. Shiu, and M. D. Lehti-Shiu. High-throughput measurement of plant fitness traits with an object detection method using faster r-cnn. *New Phytologist*, 234(4):1521–1533, 2022. 2

[47] A. P. Whan, A. B. Smith, C. R. Cavanagh, J.-P. F. Ral, L. M. Shaw, C. A. Howitt, and L. Bischof. GrainScan: A low cost, fast method for grain size and colour measurements. *Plant Methods*, 10(1):23, 2014. 1

[48] J. Wilhelm, T. Wojciechowski, J. A. Postma, D. Jollet, K. Heinz, V. Böckem, and M. Müller-Linow. Assessing the storage root development of cassava with a new analysis tool. *Plant Phenomics*, 2022, 2022. 1

[49] P. Zapotoczny. Discrimination of wheat grain varieties using image analysis and neural networks. Part I. Single kernel texture. *Journal of Cereal Science*, 54(1):60–68, July 2011. 1

[50] C. Zhang, W. A. Craine, R. J. McGee, G. J. Vandemark, J. B. Davis, J. Brown, S. H. Hulbert, and S. Sankaran. Image-based phenotyping of flowering intensity in cool-season crops. *Sensors*, 20(5):1450, 2020. 1