High-throughput 3D shape completion of potato tubers on a harvester*

Pieter M. Blok[a,*], Federico Magistri[b], Cyrill Stachniss[b], Haozhou Wang[a], James Burridge[a] and Wei Guo[a]

[a]Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Midori-cho, Nishitokyo-city, 188-0002, Tokyo, Japan
[b]Center for Robotics, University of Bonn, Nussallee 15, Bonn, 53115, North Rhine-Westphalia, Germany

ARTICLE INFO

Keywords:
3d shape completion
potato
deep learning
rgb-d
structure-from-motion

ABSTRACT

Potato yield is an important metric for farmers to further optimize their cultivation practices. Potato yield can be estimated on a harvester using an RGB-D camera that can estimate the three-dimensional (3D) volume of individual potato tubers. A challenge, however, is that the 3D shape derived from RGB-D images is only partially completed, underestimating the actual volume. To address this issue, we developed a 3D shape completion network, called CoRe++, which can complete the 3D shape from RGB-D images. CoRe++ is a deep learning network that consists of a convolutional encoder and a decoder. The encoder compresses RGB-D images into latent vectors that are used by the decoder to complete the 3D shape using the deep signed distance field network (DeepSDF). To evaluate our CoRe++ network, we collected partial and complete 3D point clouds of 339 potato tubers on an operational harvester in Japan. On the 1425 RGB-D images in the test set (representing 51 unique potato tubers), our network achieved a completion accuracy of 2.8 mm on average. For volumetric estimation, the root mean squared error (RMSE) was 22.6 ml, and this was better than the RMSE of the linear regression (31.1 ml) and the base model (36.9 ml). We found that the RMSE can be further reduced to 18.2 ml when performing the 3D shape completion in the center of the RGB-D image. With an average 3D shape completion time of 10 milliseconds per tuber, we can conclude that CoRe++ is both fast and accurate enough to be implemented on an operational harvester for high-throughput potato yield estimation. Our code, network weights and dataset are publicly available at https://github.com/UTokyo-FieldPhenomics-Lab/corepp.git.

1. Introduction

Potatoes (*Solanum tuberosum*) are important for global food security and can complement a cereal-based diet due to their high nutritional value and high yield (Zhang, Xu, Wu, Hu and Dai, 2017). To further optimize potato yields, farmers must have sub-field level information on field yields at a 5-10 meter scale, as this provides important insights into the marketable yield of the field, as well as to further optimize the fertilization and cultivation practices. Sub-field level potato yield monitoring can be performed by systems installed on the harvester that can automatically estimate key yield parameters such as tuber number, size, volume, and weight. A commonly used potato yield monitoring system uses load cells attached to the harvesters' conveyor belt to measure the mass of the harvested produce in real time (Zamani, Ghoşamiparashkohi, Faghavi and Ghezavati, 2014; Kabir, Myat Swe, Kim, Chung, Jeong and Lee, 2018). Such weighing systems are simple to use and maintain, but a disadvantage of these systems is that tare (such as stones, soil and crop residue) is also included in the weight measurement, making the tuber yield profile unobtainable.

A more promising alternative is to use a camera system that can visually distinguish potato tubers from tare, allowing the system to count and measure only the potato tubers. Currently, at least 14 of such camera systems have been presented in scientific literature: Noordam, Otten, Timmermans and van Zwol (2000); Hofstee and Molema (2003); ElMasry, Cubero, Moltó and Blasco (2012); Razmjooy, Mousavi and Soleymani (2012); Lee, Kim, Lee and Shin (2018); Long, Wang, Zhai, Wu, Li, Sun and Su (2018); Si, Sankaran, Knowles and Pavek (2018); Su, Kondo, Li, Sun, Al Riza and Habaragamuwa (2018); Pandey, Kumar and Pandey (2019); Cai, Jin, Xu and Yang (2020); Lee and Shin (2020); Dolata, Wróblewski, Mrzygłód and Reiner (2021); Huynh, TonThat and Dao (2022); Jang, Moon, Kim and Lee (2023). Most of these systems (11 of the 14) used a single red, green and blue (RGB) color camera to detect the potato tubers and then estimate their yield parameters using offline calibrated pixel-to-world conversion factors. While this method proved effective in calibrated laboratory setups, it was also found to have limited accuracy for potatoes that are occluded. Occlusion is a common phenomena on the conveyor belt of an operational harvester.

More recent studies have focused on extending the traditional RGB camera with additional three-dimensional (3D) vision. With 3D vision, it is possible to create partial or complete 3D shapes of potato tubers, allowing for better estimation of tuber yield under the challenging conditions of occlusion. Cai et al. (2020) extended an RGB camera with a laser line triangulation method to capture the complete 3D shape of potato tubers. Although their system enabled very accurate volumetric estimates of potato tubers in a laboratory setup, the image acquisition method was unfortunately too time-consuming to be applied on an operational harvester. RGB cameras with embedded depth sensing abilities, known

*This study is partially supported by the Sarabetsu Village "Endowed Chair for Field Phenomics" project in Hokkaido, Japan.

*Corresponding author: pieter.blok@fieldphenomics.com (P.M. Blok).
ORCID(s): 0000-0001-9535-5354 (P.M. Blok); 0000-0003-2815-5760 (F. Magistri); 0000-0003-1173-6972 (C. Stachniss); 0000-0001-6135-402X (H. Wang); 0000-0002-2194-9894 (J. Burridge); 0000-0002-3017-5464 (W. Guo)

as RGB-D cameras, allow much higher throughput than laser triangulation methods. RGB-D cameras, such as the ones used by Long et al. (2018) and Su et al. (2018), are therefore more promising for use on an operational harvester.

A current limitation with RGB-D cameras is that they can only produce a partial 3D shape of the potato tuber, which can lead to an underestimation of the actual size, volume or weight. Combining multiple RGB-D cameras can potentially reduce the effect of this problem, but this will not fully solve it and make the overall system complex and more expensive. Therefore, it is more desirable and economically feasible to use a single RGB-D camera, and then estimate the complete 3D shape with dedicated software. There are numerous examples in the scientific literature of using 3D shape completion software to estimate complete shapes from partial shapes. Most of the current shape completion methods use deep learning techniques based on multi-layered perceptrons, graph-based convolutional neural networks and encoder-decoder networks to complete the 3D shape from partially completed point clouds (Fei, Yang, Chen, Li, Li, Ma, Hu and Ma, 2022). Also in the agricultural domain, there are a few studies on 3D shape completion: one on the completion of plant leaves (Chen, Liu, Wang, Wang, Gong, Li and Lan, 2023), one on the completion of trees (Xu, Chen and Jing, 2023), and five on the completion of fruits (Ge, Xiong and From, 2020; Magistri, Marks, Nagulavancha, Vizzo, Läebe, Behley, Halstead, McCool and Stachniss, 2022; Magistri, Marcuzzi, Marks, Sodano, Behley and Stachniss, 2024; Marangoz, Zaenker, Menon and Bennewitz, 2022; Pan, Magistri, Läbe, Marks, Smitt, McCool, Behley and Stachniss, 2023).

When assessing the applicability of 3D shape completion methods for potato yield estimation, it is crucial to consider their processing speed. Ideally, the shape completion method should be fast enough to process all the potato tubers that move over the conveyor belt during harvest, because this will provide the most complete and accurate yield estimate possible. Since roughly more than 200,000 potato tubers are harvested per hectare, the 3D shape completion method must complete processing in a matter of milliseconds. Of the studies mentioned above, only the method of Magistri et al. (2022) has the potential to do so. Magistri's method, called CoRe (Completion and Reconstruction), can complete 3D shapes in 4 milliseconds (tested on sweet peppers and strawberries). The novelty of CoRe is the addition of a convolutional encoder to the deep signed distance function decoder (DeepSDF, Park, Florence, Straub, Newcombe and Lovegrove, 2019). CoRe's convolutional encoder compresses RGB-D images into a latent vector, which is a compact representation of the shape's geometry. The latent vector from the encoder is then used along with the 3D coordinates of the point cloud as input for DeepSDF to reconstruct the complete 3D shape. Because the latent vector is predicted by the convolutional encoder, there is no need for DeepSDF's original time-consuming latent vector optimization, making the entire 3D shape completion method significantly faster.

Although the work by Magistri et al. (2022) shows the potential for high-throughput 3D shape completion, there are still unanswered research questions. First, what is the optimal size of the latent vector for implicitly learning the shape of the potato tubers? Second, given that the potato tubers are transported on the conveyor belt and thus move from bottom to top in the image, what is the best image location to perform the 3D shape completion? Third, and most important, is the 3D shape completion method able to quickly and accurately estimate the volume of fast moving potato tubers on an operational harvester?

The main contribution of this paper is the extension of the work by Magistri et al. (2022) to answer these research questions. We have conducted a systematic study and analysis on how to optimize this 3D shape completion method. Our optimization has led to several improvements to CoRe's original convolution encoder: an improved data preprocessing, geometric and color data augmentations, an updated loss function, an updated neural network architecture, and a graphical processing unit (GPU)-based 3D mesh generation that improves and speeds up the 3D shape completion.

This paper presents the research and development of a 3D shape completion network for estimating the volume of individual potato tubers on an operational harvester. Our research novelties are four-fold. First, we optimized CoRe's original convolutional encoder for faster and more accurate 3D shape completion of potato tubers on an operational harvester. Second, we performed a systematic analysis of the effect of the latent size, image analysis region, potato size, and potato cultivar on the performance of the 3D shape completion of potato tubers. Third, we conducted two ablation studies on the impact of our new additions to the overall performance. Fourth, we publicly released our 3D dataset with partially and fully completed point clouds of potato tubers collected on an operational harvester in dirty and cluttered circumstances.

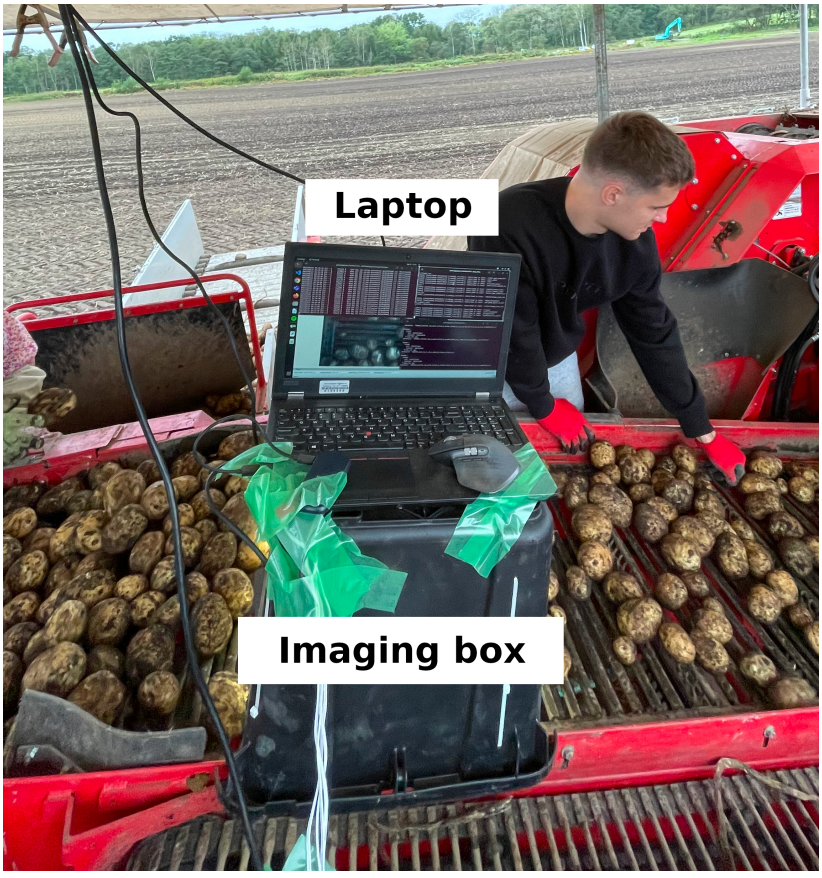
2. Materials and methods

2.1. Dataset

2.1.1. Imaging system

We installed an imaging system above the conveyor belt of a single-row potato harvester (Toyonoki Top-1, Figure 1a). The imaging system was a black plastic box in which four light emitting diode (LED) strips were installed to provide light (Figure 1b). Reflective curtains on the sides of the plastic box helped to diffuse the light (Figure 1c). In the center of the box, an RGB-D camera (Intel Realsense D405) was installed. The distance between the RGB-D camera and the conveyor belt was approximately 0.33 m. At this distance, the camera's field of view was 0.64 m (width) by 0.39 m (height). The RGB-D camera was connected to a laptop computer (Lenovo P51) that used the Robot Operating System 2 (ROS2, version: Humble Hawksbill) to collect color and depth images with 30 frames per second in ROS2 bag files.

(a)



(b)

(c)

Figure 1: (a) Overview of the imaging system installed on a potato harvester in Sarabetsu, Japan. (b) Close-up photo of the imaging system on the harvester. (c) Inside the imaging box, an RGB-D camera (Intel Realsense D405) was installed. Four LED strips provided the necessary illumination inside the box. The sides of box were covered with a reflective curtain that provided diffuse lighting conditions.

2.1.2. Image collection on the harvester

RGB-D images were collected from 12 rows in a potato field in Sarabetsu, Japan (latitude: 42.610316, longitude: 143.156753). The row spacing in this field was 0.75 m and the potato cultivar was Sayaka. For each row, a separate ROS2 bag file was recorded and stored on the laptop. To increase diversity in potato sizes and shapes, we also collected data by running the harvester in the barn of the farm

(a) (b) (c)

Figure 2: Our image collection method involved marking potato tubers with a colored thumbtack so that the tuber could be easily identified in the image and easily collected after image acquisition. (a) In this example, the tuber with the red thumbtack was first fully photographed when it entered the image acquisition area of the RGB-D camera in the bottom of the image. (b) This is the 15th captured image in which the marked potato is in the center of the image. (c) This is the 29th and last captured frame in which the marked tuber was fully photographed.

and then dumping boxes of two different potato cultivars on the conveyor belt. Although this mimicked the density of potato tubers on the conveyor belt in the field, there was no tare in the boxes, making the detection task easier. The barn experiment was conducted with six boxes of potatoes, each weighing between eight and ten kilograms. The six boxes were divided into three boxes with potatoes from the Kitahime cultivar and three boxes from the Corolle cultivar. A separate ROS2 bag file was recorded for each box and stored on the laptop.

While running the harvester in the field or in the barn, we collected images of 339 potato tubers of different sizes and shapes to test our 3D shape completion method. Our collection method consisted of the following procedure: one person standing in front of the imaging box selected a potato tuber of a random size and shape and then inserted a colored thumbtack into the potato. The thumbtack was inserted into the potato such that it was visible in the RGB image when the tuber passed the image acquisition region of the RGB-D camera. Given the speed of the conveyor belt, we were able to capture between 20 and 30 images of the marked potato tuber as it moved under the camera along with the conveyor belt (Figure 2). After the image acquisition, another person standing behind the imaging box collected the marked potato by hand. The marked potato was then placed in a bucket that was later brought to the barn for 3D reconstruction.

2.1.3. 3D reconstruction of the collected potato tubers

To obtain the complete 3D shape of the collected potato tubers, we have set up an imaging system in the barn (Figure 3a). Our imaging system consisted of three Canon X7 DSLR cameras, a turntable, four auto-detectable marker stands (Fig. 3b), and a photo studio with LED illumination.

Prior to taking the images, each collected potato tuber was pierced with a narrow threaded bolt attached to a white wooden board (Figure 3b). This piercing allowed the tuber to be photographed from three camera perspectives while being held off the ground, allowing almost the entire longitudinal section of the tuber to be photographed at once. To capture the different sides of the tuber, the turntable was set to automatically rotate 15 degrees and then stop for two seconds to give the three cameras time to acquire the images. This was repeated until the tuber was photographed from all sides. The image acquisition was accomplished with a commercial shutter controller (Esper TriggerBox) that released an electronic trigger to the three cameras at the same time. The images were directly transferred to the connected laptop computer to store them for later 3D reconstruction (Figure 3c). The camera parameter settings, image transferring, and image renaming were accomplished with DigiCamControl software, of which its details are shown in Figure 3d.

After the image acquisition, the images were processed to extract the region of the potato tuber. This region extraction was accomplished by applying a threshold on the CIELAB color space to filter out the white- and black-colored background (Figure 3e) (the threshold was set to larger than 15 for the A channel and B channel). The resulting mask was refined by using the pre-trained CascadePSP (Cheng, Chung, Tai and Tang, 2020) deep learning network (Figure 3g&h).

After generating the masks of all tubers, we implemented an automatic batch processing workflow based on the Metashape API for 3D reconstruction (Figure 3i). First, the images and corresponding masks of each potato tuber were grouped for each camera and its corresponding view angles. Then, the control point markers in the images were automatically detected, the scale bars were automatically imported, and world coordinates were automatically assigned to each potato tuber. Afterwards, the images of the different cameras were aligned, and the corresponding key points were generated. These aligned images served as input for the 3D reconstruction, which resulted the 3D mesh model, including its colored model textures. After 3D reconstruction, small distortions, such as 3D meshes belonging to the white wooden board were removed. Then, the filtered meshes were double-checked by another researcher in CloudCompare who removed small disconnected components if necessary. Figure 4 illustrates a filtered 3D mesh with colored textures of a potato tuber.

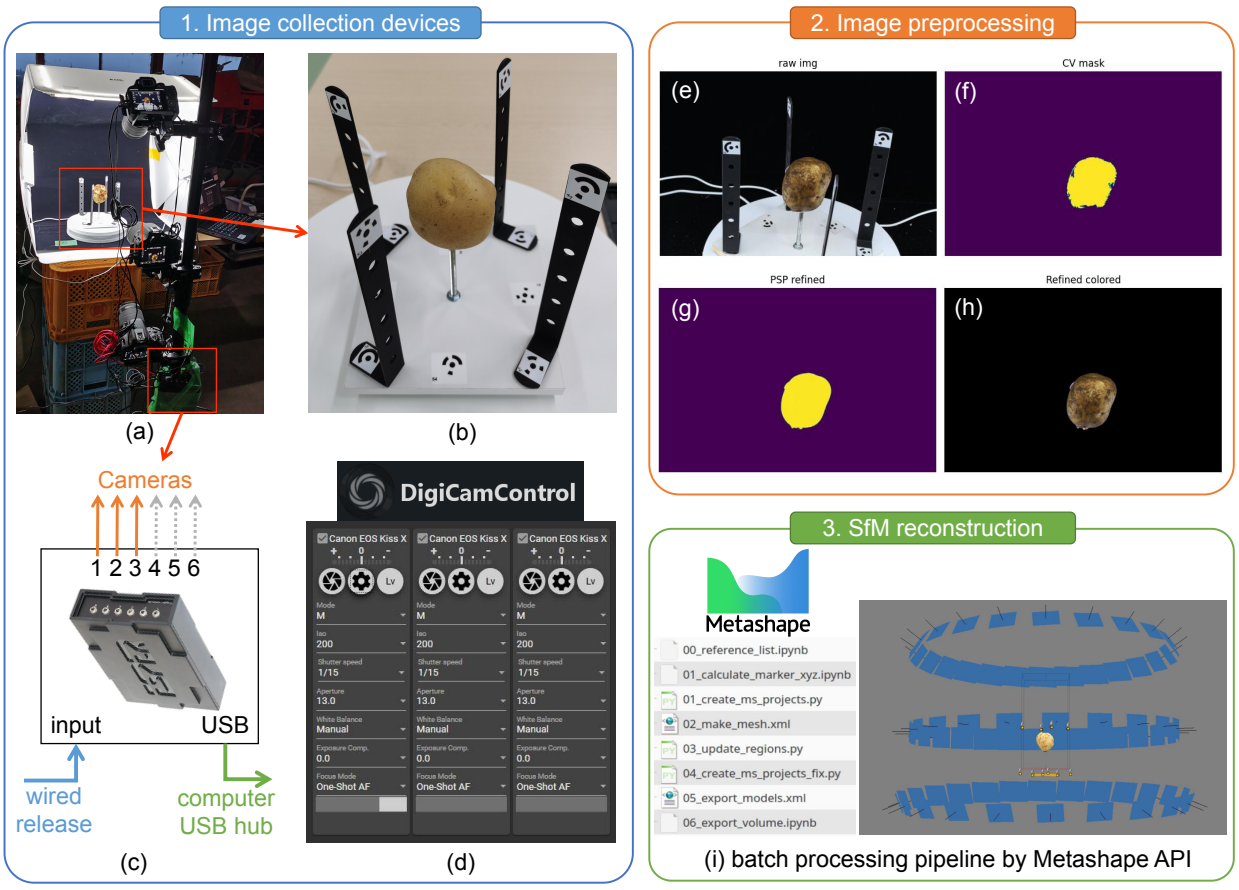


Figure 3: The workflow of our 3D reconstruction included three steps: (1) image collection, (2) image preprocessing, (3) 3D reconstruction with Structure-from-Motion (SfM). (a) Our image acquisition hardware consisted of three Canon DSLR cameras attached to a frame, a turntable, four auto-detectable marker stands, and a photo studio with LED illumination. (b) Before image acquisition, a threaded bolt was inserted into the potato tuber such that it was lifted off the base. This allowed the three cameras to photograph an entire longitudinal section of the tuber at once. (c) A multi-camera control trigger box was used to acquire images from the three cameras at the same time. (d) DigiCamControl software was used to adjust the cameras' parameters. (e-h) After image acquisition, the region of the potato tuber was extracted through CIELAB thresholding (f) and deep learning refinement (g). (i) Metashape batch processing was used to reconstruct the 3D shape of the potato tubers.

(a) (b) (c) (d)

Figure 4: 3D colored mesh of a potato tuber produced by our 3D reconstruction pipeline. (a) front view, (b) right side view, (c) back view, (d) left side view.

2.1.4. Dataset splits

After the 3D reconstruction of the 339 collected potato tubers, we split the dataset into a train, validation and test set. The split was made in such a way that the three sets contained a representative portion of different sizes, shapes and cultivars (Figure 5). Our split ratio was 70% for the train set (237 potato tubers), 15% for the validation set (51 potato tubers) and 15% for the independent test set (51 potato tubers).

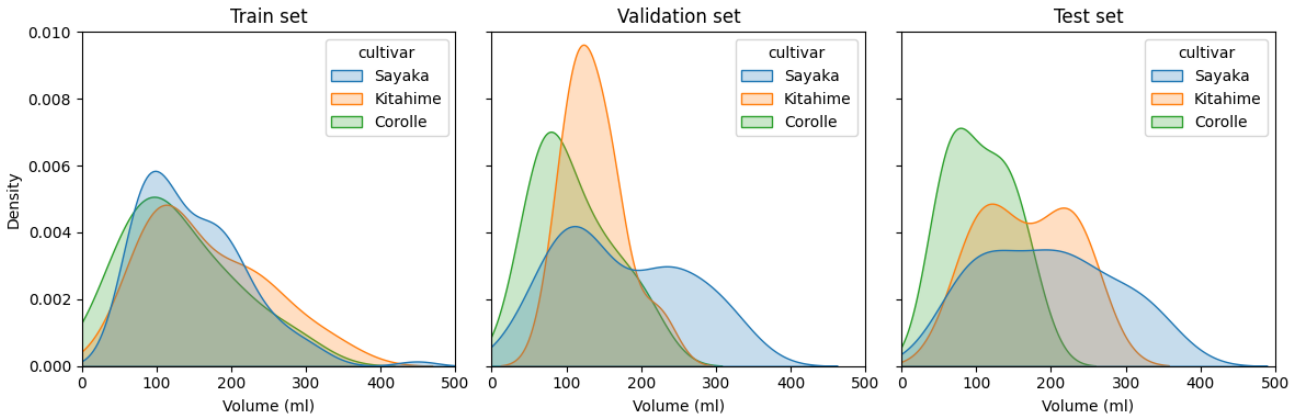


Figure 5: Kernel density estimate plots for visualizing the volumetric distribution by potato cultivar in the train, validation and test set.

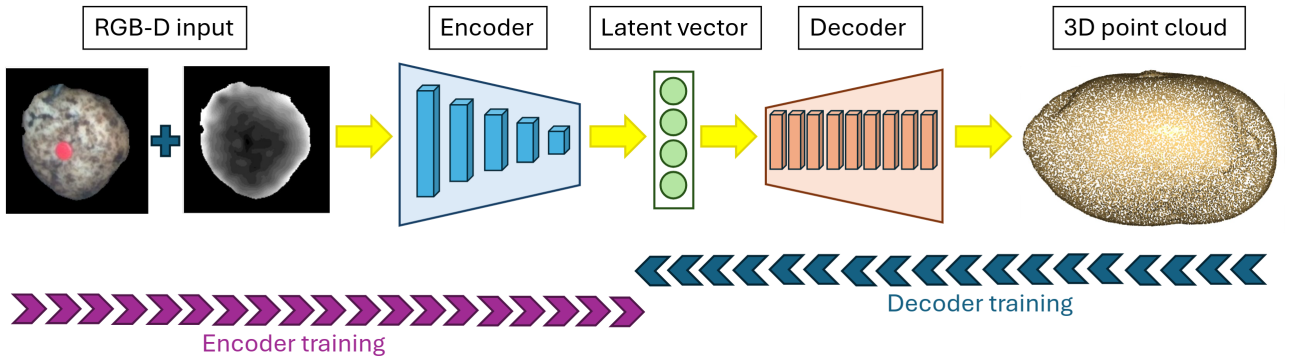


Figure 6: Schematic representation of our encoder-decoder network for 3D shape completion of potato tubers. The input data for our network was a four channel RGB-D image that was clipped and masked to the potato region and padded to a fixed dimension of 304 pixels. The input image was fed into the convolutional encoder which compressed the data into a latent vector, whose dimension was configurable. Then, the latent vector was processed by the DeepSDF decoder to reconstruct the complete 3D point cloud.

Because each potato tuber was photographed 20 to 30 times, the total number of RGB-D images for training the encoder was 6794. The number of validation images was 1439 and the number of independent test images was 1425.

2.2. 3D shape completion network

2.2.1. Encoder and decoder architecture

Our 3D shape completion network was based on the CoRe network presented in Magistri et al. (2022). Given the many updates made to CoRe during our research, our 3D shape completion network was renamed to CoRe++. Figure 6 gives a schematic overview of CoRe++. CoRe++ consisted of a convolutional encoder and a decoder. The encoder was a small neural network consisting of seven convolutional layers, each followed by a Leaky-ReLU activation function and a Max-Pooling layer, see the details in Table 1. The position of the pooling layer was one of the changes of CoRe++ compared to CoRe (in CoRe, the pooling layer was before the activation). We hypothesized that this positional change of the pooling layer could help prevent the loss of important features. After the last convolution block, there was a flatten layer after which the flattened output tensor was fed into a fully connected layer. This layer outputted the latent vector, whose dimension was made configurable for the purpose of our study. Conceptually, the latent vector is a compressed representation of the 3D shape of the potato tuber.

The decoder of CoRe++ was the coded-shape deep signed distance function (DeepSDF, Park et al., 2019). This decoder was composed of nine fully connected layers with a feature dimension output of 512. The first fully connected layer had an input dimension of latent size + 3, which was the result of concatenating the latent vector with the 3-dimensional vector of the object's coordinates in x, y, z. After the last fully connected layer there was a ReLU activation function followed by a hyperbolic tangent function. The latter outputted values between -1 and 1. Conceptually, this output is the signed distance to the object's surface, in which positive values represent 3D points outside the object's surface and negative values represent 3D points inside the object's surface. The value's magnitude corresponded to the distances to the object's 3D surface, where values close to

Table 1
Neural network architecture of the convolutional encoder of CoRe++.

Layer	Type	Kernel size	Strides	Padding	Activation	Trainable parameters
1	Convolution	3×3	1	1	Leaky ReLU (0.2)	592
2	Max Pooling	4×4	2	1	-	0
3	Convolution	3×3	1	1	LeakyReLU(0.2)	4640
4	Max Pooling	4×4	2	1	-	0
5	Convolution	3×3	1	1	LeakyReLU(0.2)	18,496
6	Max Pooling	4×4	2	1	-	0
7	Convolution	3×3	1	1	LeakyReLU(0.2)	73,856
8	Max Pooling	4×4	2	1	-	0
9	Convolution	3×3	1	1	LeakyReLU(0.2)	295,168
10	Max Pooling	4×4	2	1	-	0
11	Convolution	3×3	1	1	LeakyReLU(0.2)	1,180,160
12	Max Pooling	4×4	2	1	-	0
13	Convolution	3×3	1	1	LeakyReLU(0.2)	4,719,616
14	Max Pooling	4×4	2	1	-	0
15	Flatten	-	-	-	-	0
16	Fully Connected	-	-	-	-	131,104
					Total:	**6,423,632**

zero approximated the object's 3D shape. The concept of signed distances is visualized in Figure 7.

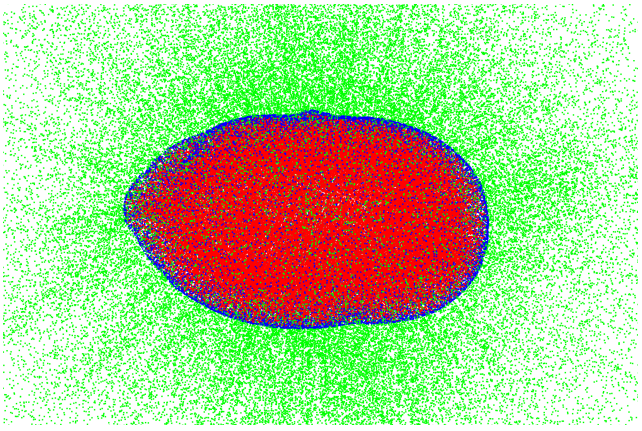


Figure 7: Visualization of the signed distances as target values for DeepSDF to learn the 3D shape of potato tubers. The red colored points are negative distance values representing the 3D points inside the tuber's surface. The green colored points are positive distance values representing the 3D points outside the tuber's surface. The blue colored points have a signed distance of zero and represent the points on the target 3D shape for DeepSDF to learn.

2.2.2. Data input, preprocessing, augmentation, and postprocessing

The input for the CoRe++ network was a four-channel RGB-D image that was masked to the potato region and clipped to a fixed-sized image of 304 x 304 pixels (Figure 6). The image dimension was chosen such that the largest tuber in our dataset would completely fit in the image. The input boxes and masks were obtained after manually annotating the potato tuber regions with the LabelMe annotation software. Note that this clipping and masking can also be achieved with an object detection or instance segmentation model.

Two new data preprocessing techniques were added to CoRe++. The first was a depth pixel filtering algorithm, which was extracted from Blok, van Henten, van Evert and Kootstra (2021), that removed all depth pixels that were further away from the depth values of the majority of the pixels in the masked depth image. This majority of pixels was assumed to represent the potato region, and helped to remove depth outliers in the depth image. The second preprocessing technique was a normalization of the remaining depth pixels between the minimum (230 mm) and maximum distance value (350 mm) between the RGB-D camera and the potato tubers on the conveyor belt. This normalization helped to increase contrast in the depth image, which potentially leads to better model convergence.

In addition to these new data preprocessing techniques, we also added geometric and color data augmentations to the encoder training procedure. The geometric augmentations involved a random image rotation to a maximum of 45 degrees, a random horizontal flipping of the image and a random vertical flipping of the image. The geometric data augmentations were applied to both the RGB image and the depth image. The color augmentations were only applied to the RGB image, and involved a random change of the brightness, saturation and hue of the image. The parameters (0.5 for brightness, 0.5 for saturation and -0.1 to 0.1 for hue) were chosen such that the augmented image looked different but visually realistic compared to the original image.

The data augmentation for training the DeepSDF decoder involved a random scaling of the original 3D shape

with scaling parameters between 0.5 and 2.0, meaning that the original shape was scaled between half of its original size and double of its size. The second data augmentation involved a rotation of the 3D shape around the Z axis by a rotation value between 0.0 and 30.0 degrees. The last data augmentation was a random shear of the 3D shape of maximally 0.5 in the X direction. For each original 3D shape, 10 augmented 3D shapes were included into the train set.

As Figure 6 indicates, the output of the CoRe++ network is a completed 3D point cloud. Since it is not possible to calculate volume directly from a 3D point cloud, we added a 3D data postprocessing procedure that converted the 3D point cloud into a watertight 3D mesh from which the volume can be estimated. Our 3D mesh generation procedure consisted of selecting the predicted signed distance values less than or equal to 0.0. These values represented the 3D points on the surface and inside the potato tuber, refer to the blue- and red-colored points in Figure 7. From the selected points, a 3D convex hull shape was extracted using Open3D's GPU-accelerated hull function. The convex hull was then smoothed with a triangle-based algorithm (Loop, 1987) that divided each triangle of the hull into four smaller triangles. What followed were noise suppression procedures and an iterative voxel-based downsampling of the generated hull in case it was not watertight. The latter procedure guaranteed that each produced 3D mesh was watertight so that the volume could be calculated.

2.2.3. Training procedure

The encoder and decoder were trained in the opposite order: first the DeepSDF decoder was trained on the complete 3D shapes in order to optimize the latent vectors. Then, the encoder was trained to fit the preprocessed RGB-D images to the target latent vectors. In Figure 6, these two training procedures are visualized by the blue colored and purple colored arrows, respectively.

At the start of the DeepSDF training, a randomly initialized latent vector was assigned to each data point. These latent vectors were then optimized along with the weights of the decoder using standard backpropagation. In our research, the DeepSDF decoder was trained for 1001 epochs with a step-based learning rate scheduler that started at a learning rate of $5{\cdot}10^{-4}$. For every 300 epochs, the learning rate was halved. The decoder was trained with the Adaptive Moment Estimation (Adam) optimizer. We used the L1 loss function during training, because this loss function serves the purpose of minimizing the sum of all distance differences between the target 3D shape and the predicted 3D shape. For every 10 epochs, a weight file and a latent vector code were automatically saved. After training, the weight file with the best 3D reconstruction on the shapes of the validation set was selected. This selected weight file and the corresponding latent vector code were used as targets for training the encoder. The metric for determining the best 3D reconstruction of the DeepSDF decoder was the Chamfer distance. The Chamfer distance (d_{CD}) represents the sum of the average closest distance from points in the ground truth 3D shape ($\mathcal{G}$) to the points in the reconstructed 3D shape ($\mathcal{S}$), and vice versa (Equation 1). The lower the Chamfer distance, the better the 3D reconstruction.

$$d_{CD}(\mathcal{G},\mathcal{S}) = \frac{1}{|\mathcal{G}|}\sum_{x\in\mathcal{G}}\min_{y\in\mathcal{S}}\|x-y\|_2^2 + \frac{1}{|\mathcal{S}|}\sum_{y\in\mathcal{S}}\min_{x\in\mathcal{G}}\|y-x\|_2^2 \tag{1}$$

where $|\mathcal{G}|$ and $|\mathcal{S}|$ are the numbers of points in $\mathcal{G}$ and $\mathcal{S}$, respectively

The weights of the encoder were randomly initialized before training started. During encoder training, the weights of the decoder were frozen. The encoder was trained for 100 epochs. The initial learning rate was $1{\cdot}10^{-4}$, and this value was gradually decreased by an exponential learning rate scheduler of 97%. The encoder was also trained with the Adam optimizer. We used the mean squared error loss function (MSE) to minimize the squares of the differences between the target latent vector and the predicted latent vector. We chose this loss function because it penalizes larger differences between the target and predicted latent vector more heavily than CoRe's default L1 loss function. Besides the MSE loss function, we used the contrastive loss function by Magistri et al. (2022). The rationale is that contrastive loss encourages the encoder to learn latent vectors that are well-separated in the latent space for the different potato tubers. Contrastive loss also enforces the encoder to learn latent vectors closer in the latent space for the images belonging to the same potato tuber. The contrastive loss function is summarized in Equation 2, where N is the total number of potato tubers, $\mathbf{z}_i$ and $\mathbf{z}_j$ represent the latent vector of instances i and j, respectively, y_i and y_j denote the potato tuber identifiers of instances i and j, respectively, δ_{rep} is the margin parameter controlling the minimum separation between representations of instances with different tuber identifiers (in our research δ_{rep} was set to 0.5), $\|\cdot\|$ denotes the Euclidean distance between two latent vectors, $\|\cdot\|_+$ is the positive part of the argument, ensuring that only positive differences contribute to the loss.

$$L_{\text{c}} = \sum_{i=1}^{N}\sum_{j=1}^{N}\begin{cases}\|\mathbf{z}_i-\mathbf{z}_j\|_+, & \text{if } y_i = y_j\\ \max\{0,\delta_{\text{rep}}-\|\mathbf{z}_i-\mathbf{z}_j\|\}, & \text{if } y_i \neq y_j\end{cases} \tag{2}$$

The combined loss function for training the encoder is summarized in Equation 3. The loss contribution values were set to 1.0 for w_{mse}, and 0.05 for w_c. These values were optimized in a pre-comparative experiment.

$$\mathcal{L} = w_{mse}\cdot\mathcal{L}_{mse} + w_c\cdot\mathcal{L}_c \tag{3}$$

For every epoch the encoder was inferred on the validation set with the decoder activated. The encoder weights with the lowest root mean squared error (RMSE, Equation 4)

on the volume were stored for final evaluation on the independent test set. We chose the RMSE metric, because it was the most common metric in the scientific literature for evaluating the volumetric estimate of potato tubers. The ground truth volume (V, Equation 4) was extracted from the reconstructed 3D mesh (Section 2.1.3). This volume estimate was more accurate than the volume estimate of the ground truth method with water displacement, as the latter had an accuracy of 10 ml at best and was more prone to human reading and writing errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{V} - V^2} \tag{4}$$

where $\hat{V}$ and V are the estimated volume and the ground truth volume, respectively. n is the number of potato tubers in the dataset.

2.3. Evaluation

The Chamfer distance (Equation 1, Section 2.2.3) was used as the performance metric for evaluating the 3D point cloud completion. Besides the Chamfer distance, we also calculated the precision, recall and the F-score. The precision (p) was the percentage of reconstructed points within a certain distance (d) to the ground truth point cloud (Equation 5). As such, it represented the accuracy of the 3D reconstruction. The recall (r) was the percentage of ground truth points within a certain distance (d) to the reconstructed point cloud (Equation 6). The recall represented the completeness of the 3D reconstruction. For our evaluation, we used a distance (d) threshold of 5.0 mm, as this value allowed to compare our results one-on-one with those of Magistri et al. (2022). After calculating the precision and recall, the F-score was obtained (Equation 7). The F-score was the harmonic mean between the precision and the recall, and it represented the percentage of 3D points that were correctly reconstructed.

$$precision(d) = \frac{100}{|S|}\sum_{y\in S}\left[\!\left[\min_{x\in\mathcal{G}}\|y-x\| < d\right]\!\right] \tag{5}$$

$$recall(d) = \frac{100}{|\mathcal{G}|}\sum_{x\in\mathcal{G}}\left[\!\left[\min_{y\in S}\|x-y\| < d\right]\!\right] \tag{6}$$

$$f\text{-}score(d) = \frac{2\cdot p(d)\cdot r(d)}{p(d)+r(d)} \tag{7}$$

For evaluating the volumetric estimate we used the RMSE metric, as presented in Equation 4. To evaluate the processing speed, we calculated the average 3D shape completion time on the test images. This analysis was performed with a Lenovo Legion Pro 7 16IRX8H laptop with a NVIDIA GeForce RTX 4090 Laptop GPU.

2.4. Experiments

2.4.1. The effect of the latent size on the 3D shape completion result

For 3D shape completion, it is important that the encoder-compressed latent vector contains high-level distinguishable 3D features that generalize well on new and untrained shapes. Because the latent vector plays such an important role in an encoder-decoder network like ours, we set up an experiment that tested what was the optimal size for the latent vector for completing the 3D shape of potato tubers. We investigated six sizes: 8, 16, 32, 64, 128, and 256. Latent size 32 was used in the original study by Magistri et al. (2022). Our hypothesis was that the smaller sizes, such as 8 and 16, had insufficient capacity to compress the 3D shape, leading to suboptimal generalization. Also, we hypothesized that the larger sizes, such as 128 and 256, were probably too large resulting in insufficient compression to generalizable 3D shape features. Hence, we hypothesized that the medium sizes, 32 and 64, were probably more optimal for the compression and thus generalization of 3D shapes.

To better understand the optimal size of the latent vector, we conducted an additional experiment with our CoRe++ network. This experiment consisted of interpolating the values of the latent vector of the smallest potato tuber and the largest in the test set. By interpolating the values between these two latent vectors, we were able to better understand the diversity of the generated 3D shapes and sizes in the corresponding latent space of the six tested latent sizes. Ideally, the latent space is constructed so that after interpolation, the generated 3D shapes are realistic in shape and have a sequential build-up in size, as the interpolation was performed between the smallest and largest tuber.

The experiment was tested with 3 networks: the DeepSDF decoder-only architecture without the encoder, the original CoRe network by Magistri et al. (2022), and our CoRe++ network. The DeepSDF decoder-only architecture was tested to better understand where the largest effect of latent size was: in the encoder or in the decoder. Testing the original CoRe network by Magistri et al. (2022) provided insight into a potential trend commonly shared between CoRe and CoRe++. It also gave a more detailed overview on the potential improvement of CoRe++ over CoRe.

The performance parameters were the ones listed in Section 2.3. Our evaluation was based on two requirements: accuracy and analysis speed. The accuracy requirement was met if the RMSE on the volume was lower than that of a standard linear regression model. Input to that linear regression model were the length, width and depth estimates of the partially completed point cloud (Figure 8). This point cloud was obtained by converting the filtered RGB-D image (Figure 6) into 3D points in Open3D software using the intrinsic camera parameters of the RGB-D camera. From the generated point cloud, the oriented 3D bounding box was obtained from which the length, width and depth dimensions were extracted as input parameters for training the linear regression model. The linear regression model was

trained on the same train and validation set and tested on the same independent test set (Section 2.1.4). The trained linear regression model resulted a RMSE of 31.1 ml on the volumetric estimate of the test images. As such, if the 3D shape completion method had a RMSE lower than 31.1 ml, then the accuracy requirement was met.

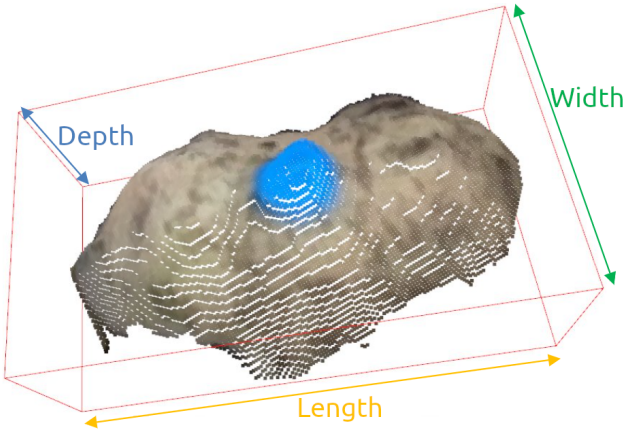


Figure 8: To determine whether the accuracy requirement was met, we trained a linear regression model on the length, width and depth estimates of the partially completed point cloud. The dimension estimates were obtained from the oriented 3D bounding box which is shown in red.

The analysis speed requirement was met if the 3D shape completion was finished in less than 16 milliseconds (ms) per potato tuber. This analysis time was calculated based on the highest number of potato tubers present in a single image in our dataset: 59. The average throughput time of the potato tubers on the harvester was 0.95 seconds, and this value was derived by dividing the average number of 28.5 frames per tuber (9658/339, Section 2.1.4) by the camera acquisition rate of 30 frames per second. With the maximum number of tubers in a single image and the average throughput time, we estimated that the 3D shape completion method should analyze up to 62 potato tubers per second (59/0.95). This equaled to 16 ms per tuber (1000/62).

2.4.2. The effect of the potato size, potato cultivar, and image analysis region on the 3D shape completion result

The goal of this experiment was to gather insights into the performance of our CoRe++ shape completion method when applied in the field. The first practical analysis was on the effect of the tuber size on the 3D shape completion result. This analysis is important because it provides information about the degree of generalizability of our method in a particular field. Under typical field conditions, a wide variety of tuber sizes move over the harvester's conveyor belt, which means that our 3D completion method should be able to generate accurate volumetric estimates regardless of the tuber size. We decided to perform the analysis on four different size classes, which were extracted from our dataset distribution (Figure 5): small tubers with a volume between 0 and 100 ml, small to medium tubers with a volume between 100 and 150 ml, medium to large tubers with a volume between 150 and 200 ml, and large tubers with a volume between 200 and 500 ml. The ranges were chosen so that a relatively balanced number of tubers ended up in the four classes.

The second practical analysis was on the effect of the potato cultivar on the 3D shape completion result. This analysis provided insights into the degree of generalizability of the 3D completion method at the farm level, as the majority of farmers grow multiple cultivars in a season. We wanted to test if our method can provide accurate volumetric estimates regardless of the potato cultivar. We conducted the experiment with three cultivars (Corolle, Kitahime and Sayaka) that differed in shape and size. Corolle was the cultivar with the most elongated and smallest tubers. Kitahime was the cultivar with the most spherical and medium-sized tubers. Sayaka was the cultivar with the largest tubers and the greatest diversity in shape. This may be due to the fact that Sayaka was the cultivar with the most samples in our dataset. As shape metrics, we calculated the elongation factor and the concavity factor. The elongation factor was obtained by dividing the longest dimension of the 3D bounding box of the ground truth mesh by the shortest dimension. Spherical shapes have an elongation factor close to 1 and elongated shapes have an elongation factor closer to 2. The concavity factor was obtained by calculating the Chamfer distance between the original ground truth 3D mesh and the convex hull of that mesh. Higher Chamfer distances mean that there are more valleys or concave regions on the surface of the potato tuber.

The final practical analysis was on the effect of the image analysis region on the 3D shape completion result. This analysis provided insight into the RGB-D image region where ideally the 3D shape completion should be performed. This analysis was necessary because it is computationally intensive to analyze all RGB-D frames for each tuber. We intended to identify an area in the RGB-D image where the RMSEs for the volumetric estimate are lowest on average. That particular area will be most promising for performing the high-throughput 3D shape completion. In our analysis, the RMSEs were summarized for thirteen horizontal regions in the RGB-D image (Table 2).

2.4.3. Ablation studies

Two ablation studies were performed as a final experiment. The first ablation study was conducted on the additions to CoRe++, which helped us to better understand the impact of the individual additions to the overall performance. The second ablation study was conducted on the components that were commonly shared between CoRe++ and CoRe. This ablation study helped us to better understand the impact of changes in the input data and network architecture of the convolutional encoder.

The first ablation study consisted of examining the impact of seven additions made to CoRe++. The first two

Table 2
Summary of the investigated horizontal pixel regions for analyzing the effect of the image analysis region on the 3D shape completion result.

Region	Horizontal region [pixels]
1	0 - 100
2	100 - 150
3	150 - 200
4	200 - 250
5	250 - 300
6	300 - 350
7	350 - 400
8	400 - 450
9	450 - 500
10	500 - 550
11	550 - 600
12	600 - 650
13	650 - 720

examined ablations consisted of individually deactivating the depth normalization and the depth filtering when training the convolutional encoder (both data preprocessing steps are described in Section 2.2.2). After that, we examined the impact of deactivating the data augmentation when training the convolutional encoder. The fourth ablation examined the effect of repositioning the Max-Pooling layer before the LeakyReLU activation, as was the case in the original CoRe implementation (refer to Section 2.2.1). The fifth examined ablation was changing the MSE loss function back to the L1 loss function, as was the case in CoRe (refer to Section 2.2.3). The sixth examined ablation consisted of replacing CoRe++'s method for automatically determining the best network weights (using GPU-based 3D mesh generation and the RMSE volume metric) with CoRe's original method (using marching cubes (Lorensen and Cline, 1987) mesh generation and Chamfer distance metric). The seventh examined ablation consisted of deactivating CoRe++'s 3D smoothing technique which was described in Section 2.2.3.

The second ablation study consisted of examining the effects of ten ablations in the input data and network architecture that are commonly shared between CoRe++ and CoRe. The first ablation involved training CoRe++ with a single-channel depth image instead of the original four-channel RGB-D image. This ablation gave us a better understanding of the effect of adding color channels to the input data for completing the 3D shape. The second ablation involved training CoRe++ with an RGB-D image clipped to the bounding box instead of the original mask. This ablation gave us insight into how the final application on the potato harvester should look like: one based on an object detection model or one based on an instance segmentation model. Logically, this ablation also gave insight into the future annotation effort, which would be significantly higher when using an instance segmentation model. The third ablation involved reducing the encoder's network architecture from seven to five convolutional blocks (layers 1-10 in Table 1 followed by a flatten layer and a fully connected layer). This ablation gave us insight into the effect of using an even lighter network in the situation of hardware constraints. The fourth ablation involved removing the pooling layers (the even layers in Table 1) to investigate their effect on the overall performance. The fifth ablation involved replacing CoRe++'s original LeakyReLU activation with the standard ReLU activation. The sixth ablation involved disabling the contrastive loss element in the loss function so that only the MSE loss was used during training. The seventh to the tenth ablation involved increasing or decreasing the learning rate in increments of two and five. This resulted in using a learning rate of $5{\cdot}10^{-4}$ (learning rate $\cdot$ 5), $2{\cdot}10^{-4}$ (learning rate $\cdot$ 2), $5{\cdot}10^{-5}$ (learning rate $\cdot$ 0.5) and $2{\cdot}10^{-5}$ (learning rate $\cdot$ 0.2).

3. Results

3.1. The effect of the latent size on the 3D shape completion result

Table 3 summarizes the results of the 3D shape completion for the three 3D completion methods and the six latent sizes. Interestingly, the effect of the latent size was marginal when testing the DeepSDF decoder-only network. For this network, the smallest latent size of 8 resulted in the lowest Chamfer distance and RMSE on the volume. One possible explanation for this result is that the DeepSDF decoder uses an iterative optimization process during inference to extract the best possible latent vector. This iterative process probably helped to optimize the latent vector for each potato shape in the test set, making the 3D shape completion slow (33 seconds on average), but also better optimized for each of the six tested latent sizes.

For both CoRe and CoRe++, the best latent size was 32, followed by 64. This result is consistent with our hypothesis that the medium-sized latent vectors are better for compressing the data into meaningful 3D representations. Figure 9 shows that latent sizes 32 and 64 have the most diverse 3D shapes and sizes after latent space interpolation. Compared to the other sizes, latent sizes 32 and 64 have both spherical and elongated 3D shapes, and the sizes show a proper sequential build-up. Latent size 8 and 256 are both worse in 3D shape completion performance (Table 3) and latent space interpolation (Figure 9). Latent size 8 could only produce five valid 3D shapes out of the seven interpolated latent vectors, while latent size 256 failed to produce a sequential size build-up. Latent sizes 16 and 256 produced mainly elongated 3D shapes, meaning that they had a limited ability to produce the more spherical shapes.

In terms of 3D shape completion and volumetric estimate, CoRe++ outperformed CoRe for five of the six latent sizes: 16, 32, 64, 128, 256 (Table 3). The difference between the 3D shape completion result of CoRe++ and CoRe is visualized in Figure 10. As for CoRe++, the reconstructed 3D shapes are smooth and they approximate the real shape of the potato tuber. With CoRe, the 3D shapes are rougher and

Table 3
The effect of the latent size on the 3D shape completion results with DeepSDF, CoRe, and CoRe++. The upward arrows indicate the higher the better, and the downward arrows the lower the better. The values in bold are the best performing values per 3D shape completion method. The last two columns summarize whether the predefined accuracy and speed requirements were met. Note that the accuracy requirement was met if the RMSE on the volume was less than 31.1 ml (this was the RMSE when applying linear regression). The speed requirement was met if the analysis time was less than 16 ms. Refer to Section 2.4.1.

3D shape completion method	latent size	d_{CD} [mm] ↓	f-score [%] ↑	precision [%] ↑	recall [%] ↑	RMSE [ml] ↓	time [ms] ↓	acc.	speed
DeepSDF (Park et al., 2019)	8	**1.5**	98.1	97.9	98.4	**7.2**	32636.5	✓	✗
	16	1.7	98.8	98.8	98.7	11.1	32669.0	✓	✗
	32	1.8	99.2	99.2	99.2	7.5	32580.9	✓	✗
	64	1.8	98.8	98.8	98.8	16.8	32792.3	✓	✗
	128	1.8	**99.3**	**99.4**	**99.3**	13.3	32919.1	✓	✗
	256	1.9	97.3	97.4	97.3	13.4	33003.6	✓	✗
CoRe (Magistri et al., 2022)	8	6.0	50.8	53.2	49.7	60.6	8.5	✗	✓
	16	3.7	71.8	73.8	70.0	50.3	9.3	✗	✓
	32	**3.1**	**81.4**	**81.5**	**81.5**	**36.9**	7.5	✗	✓
	64	3.4	76.1	74.1	78.6	41.6	8.3	✗	✓
	128	3.3	78.3	78.7	78.0	43.9	8.0	✗	✓
	256	5.3	58.3	58.8	58.6	90.0	**6.7**	✗	✓
CoRe++ (ours)	8	7.9	38.1	39.8	37.6	68.8	**9.1**	✗	✓
	16	4.2	65.4	66.8	64.4	44.6	13.0	✗	✓
	32	**2.8**	**85.0**	**85.2**	**85.0**	**22.6**	9.9	✓	✓
	64	2.9	83.2	83.8	82.8	28.1	9.4	✓	✓
	128	3.3	78.6	81.1	76.4	35.7	9.3	✗	✓
	256	4.5	62.5	63.7	61.8	65.0	10.2	✗	✓

spikier, leading to less resemblance to the real potato shape, larger Chamfer distances and larger volumetric errors.

Only for CoRe++ and latent sizes 32 and 64, the accuracy requirement was met as the corresponding RMSE values were less than 31.1 ml (this was the baseline value when using linear regression). The requirement for analysis speed was met for each latent size for both CoRe and CoRe++, as all of the analysis times were less than 16 ms (Table 3).

3.2. The effect of the potato size, potato cultivar, and image analysis region on the 3D shape completion result

Table 4 summarizes the effect of the potato size on the 3D shape completion result. There is a trend that the RMSE on the volumetric estimate is larger when the tuber is larger. A possible explanation is that the larger tubers are more concave (Table 4), meaning that they have more valleys and variations in their 3D curvature, making it harder to accurately reconstruct the 3D shape (especially when the concave parts are facing downward with respect to the camera's perspective). The fact that larger Chamfer distances are observed on the largest tubers supports this explanation. For the relative volumetric errors, there is an opposite trend, meaning that the largest relative volumetric errors are observed on the smaller tubers.

Table 5 summarizes the effect of the potato cultivar, and thus indirectly the shape of the potato tuber, on the 3D shape completion results. There is a trend that the Chamfer distance is lower when the tubers have a more spherical shape (i.e. elongation factors closer to 1.0). For the RMSE on the volumetric estimate, there is no clear trend regarding the tuber shape. What may explain this result is that the different cultivars had different tuber sizes: the cultivar Corolle had the relative smallest tubers and it was already shown in Table 4 that the RMSE is lowest when the tubers are smaller. The largest RMSE was observed on Sayaka, which was the cultivar with the largest tubers on average, but it was also the cultivar with the largest number of samples, meaning that there was probably more diversity in tuber shape and size.

Figure 11 visualizes the RMSE on the volumetric estimate for the thirteen different image regions. The smallest RMSE of 18.2 ml was observed in the central horizontal region of the image between 350 and 400 pixels. The largest RMSE values were observed in the lower and upper parts of the image, indicating that these regions are not recommended for performing the 3D shape completion. These results may indicate that some form of lens distortion may have occurred in the peripheral regions of the camera's field of view, leading to poorer 3D shape completion in these regions. There may also have been a higher degree of occlusion in these regions.

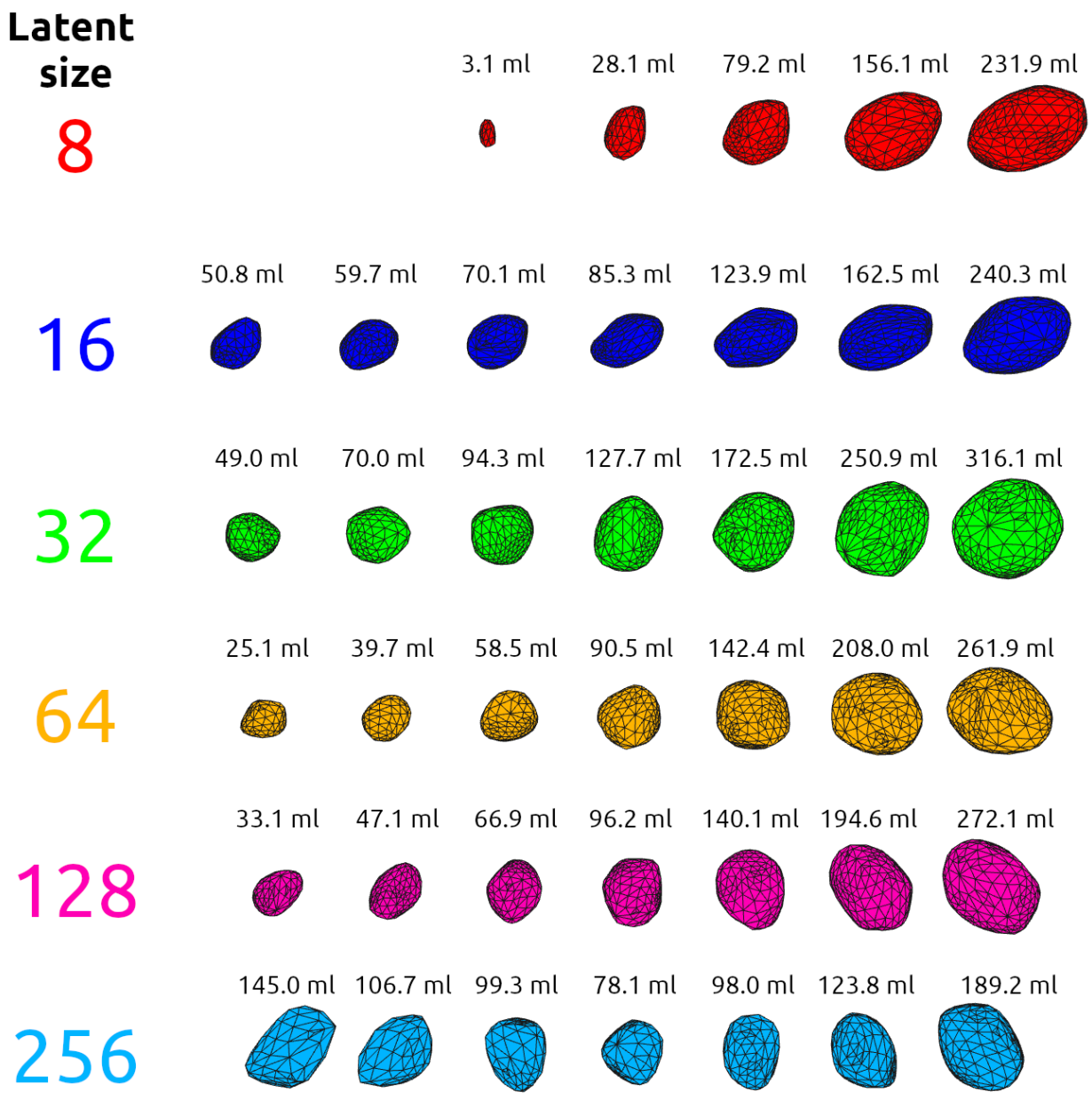
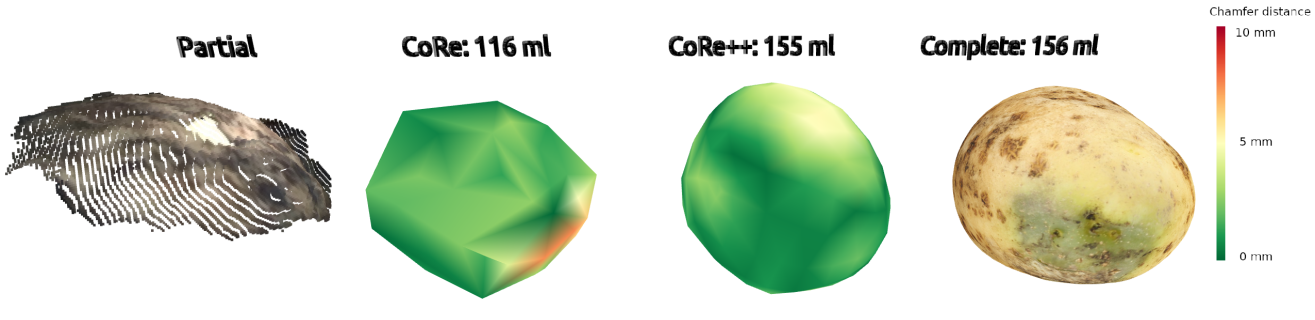


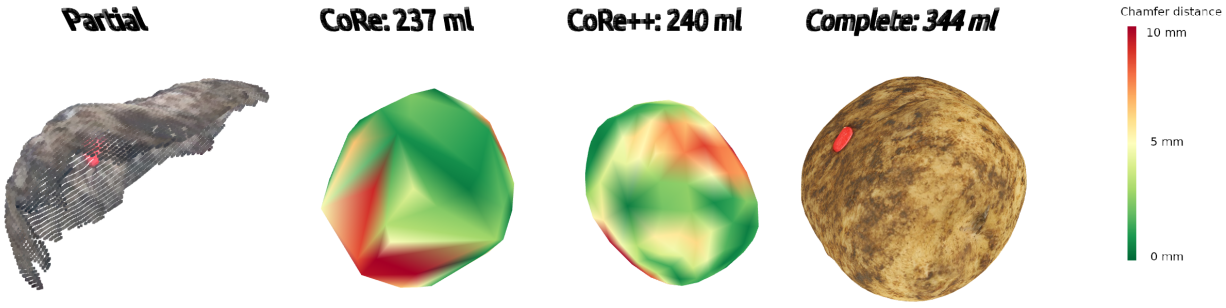
Figure 9: 3D shape completion results for the seven latent space interpolations, visualized for each of the six tested latent sizes. The values above the 3D shapes are the volumes of the generated meshes in milliliters (ml).

3.3. Ablation studies

The ablation study on CoRe++'s additions (Table 6) shows that the largest contribution to the overall performance was made by CoRe++'s validation method, which was based on GPU mesh generation and RMSE validation metric. Changing this validation method to the original validation method of CoRe with marching cubes mesh generation and Chamfer distance validation metric, led to the largest increase in both Chamfer distance (+32.1%) and RMSE on volumetric estimate (+122.1%). An obvious explanation for this result is that the final volumetric estimate benefits from having the RMSE optimized during training. Another explanation is that CoRe's original marching cubes method seems unable to accurately reconstruct the 3D shapes of the potato tubers. This may be due to the chosen grid density of the marching cubes method which was optimized for high-throughput 3D reconstruction, but possibly came at the expense of the accuracy (something that can also be observed in Figure 10). Three other additions to CoRe++ that significantly improved the overall performance were the two data preprocessing steps (highlighted by the first and second ablation) and the loss function modification to MSE (highlighted by the fifth ablation).



(a)



(b)

Figure 10: (a) CoRe++'s best 3D shape completion result was achieved on a medium-sized potato tuber. (b) CoRe++'s worst 3D completion result was achieved on a large-sized and irregularly-shaped potato tuber. In both figure (a) and (b), the partial point cloud from the Realsense D405 is visualized on the left, and this point cloud needs to be completed. The second from the left visualizes the completion result with CoRe, including the volumetric estimate in milliliters. The colors of the 3D shape represent the Chamfer distances, where the orange and red colors are above 5 millimeters and the green and yellow colors below 5 millimeters. The second from the right visualizes the completion result with CoRe++. Most right is the 3D ground truth shape from the structure-from-motion method. Note that figure (a) and (b) were not at the same scale because they were zoomed differently to reveal better details.

Table 4
3D shape completion results expressed for the four size classes. Count summarizes the total number of RGB-D frames analyzed per size class.

Volume [ml]	Count	Elongation factor	Concavity factor [mm]	d_{CD} [mm] ↓	f-score [%] ↑	precision [%] ↑	recall [%] ↑	RMSE [ml] ↓	rel. error [%] ↓
0-100	361	1.6	0.2	2.5	89.3	89.0	89.8	16.8	19.1
100-150	364	1.4	0.3	3.0	82.3	82.1	82.6	18.2	12.5
150-200	277	1.4	0.3	2.7	85.4	85.8	85.1	21.7	10.0
200-500	423	1.5	0.4	2.9	83.5	84.2	82.9	29.7	9.0

Regarding the generic ablation study (Table 6), there are a few interesting outcomes. First, the choice of the activation function has a large impact on the overall performance. The fifth ablation shows that after replacing the LeakyReLU activation with a standard ReLU, the Chamfer distance increased by 46.4% and the RMSE by 96.0%. This outcome is consistent with that of Tomar (2022), who also found that ReLU under-performs compared to the more advanced variants of ReLU when testing an autoencoder on the Global Wheat Head dataset (David, Madec, Sadeghi-Tehran, Aasen, Zheng, Liu, Kirchgessner, Ishikawa, Nagasawa, Badhon, Pozniak, de Solan, Hund, Chapman, Baret, Stavness and Guo, 2020). In our case, we think that the LeakyReLU activation resulted in a better and faster network convergence during training, due to LeakyReLU's ability to maintain non-zero gradients for negative inputs. A second interesting outcome was the relatively large impact of the RGB color channels on the overall performance. Our initial expectation was that only the depth image would contain relevant features for completing the 3D shape. There are two possible explanations for this result. First, the depth image has typically more noise than the RGB image, which may lead to less good feature extraction when only using the depth image. Second, the data augmentations for the RGB color channels were more extensive than for the depth channel, which may have resulted in a better generalization performance when using RGB-D images rather than just depth images. A third interesting outcome is that training

Table 5
3D shape completion results expressed for the three tested potato cultivars. Count summarizes the total number of RGB-D frames analyzed per cultivar.

Potato cultivar	Count	Elongation factor	Concavity factor [mm]	d_{CD} [mm] ↓	f-score [%] ↑	precision [%] ↑	recall [%] ↑	RMSE [ml] ↓	rel. error [%] ↓
Corolle	291	1.9	0.3	2.9	83.9	84.1	83.8	17.6	15.4
Sayaka	869	1.5	0.3	2.8	84.5	84.9	84.2	24.8	11.8
Kitahime	265	1.2	0.3	2.5	88.0	87.2	88.9	19.3	12.6

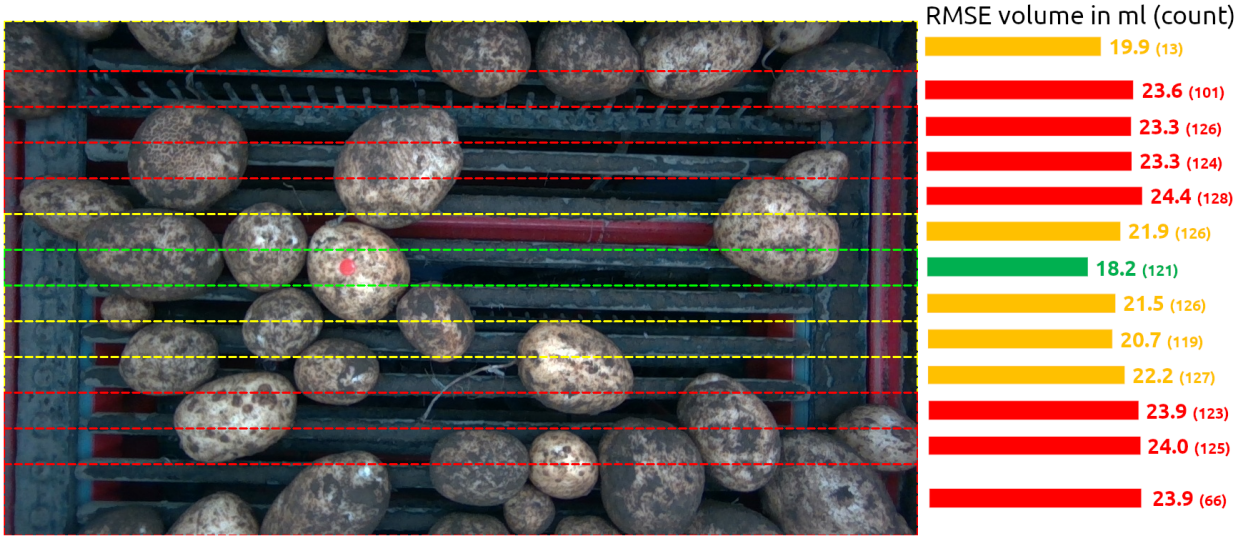


Figure 11: Root mean square errors (RMSE) visualized for thirteen horizontal image regions. The green-colored region in the center of the image between 350 and 400 pixels had the lowest RMSE of 18.2 ml.

Table 6
Performance metrics for the two ablation studies relative to the best performance of CoRe++.

Ablation	Category	$d_{CD}(mm)$ ↓ abs	$d_{CD}(mm)$ ↓ rel	f-score ↑ abs	f-score ↑ rel	RMSE (ml) ↓ abs	RMSE (ml) ↓ rel
CoRe++	Baseline	2.8	-	85.0	-	22.6	-
Ablation study on CoRe++'s additions							
No depth normalization	Data preprocessing	3.3	+17.9%	78.6	-7.5%	34.3	+51.8%
No depth filtering	Data preprocessing	3.4	+21.4%	75.9	-10.7%	41.1	+81.9%
No data augmentation	Data augmentation	2.9	+3.6%	83.9	-1.3%	32.4	+43.4%
Act-Pool → Pool-Act	Network changes	2.9	+3.6%	83.2	-2.1%	31.0	+37.2%
MSE loss → L1 loss	Loss function	3.1	+10.7%	81.3	-4.4%	31.4	+38.9%
Val → CoRe val	Train validation	3.7	**+32.1%**	72.0	**-15.3%**	50.2	**+122.1%**
Smoothing → Custom	3D postprocessing	2.8	0.0%	85.1	+0.2%	24.9	+10.2%
General ablation study							
RGB-D → D	Data preprocessing	3.5	+25.0%	75.4	-11.3%	40.5	+79.2%
Mask → Box	Data preprocessing	3.2	+14.3%	78.9	-7.2%	33.0	+46.0%
7 → 5 Conv. blocks	Network changes	3.0	+7.1%	82.3	-3.2%	32.0	+41.6%
No pooling layers	Network changes	2.9	+3.6%	83.0	-2.4%	29.7	+31.4%
LeakyReLU → ReLU	Network changes	4.1	**+46.4%**	68.1	**-19.1%**	44.3	**+96.0%**
No contrastive loss	Loss function	2.7	-3.6%	86.3	+1.5%	23.6	+4.4%
LR → LR·5	Learning rate	2.9	+3.6%	83.6	-1.6%	32.1	+42.0%
LR → LR·2	Learning rate	2.8	0.0%	85.2	+0.2%	28.8	+27.4%
LR → LR·0.5	Learning rate	3.0	+7.1%	81.9	-3.6%	31.0	+37.2%
LR → LR·0.2	Learning rate	3.3	+17.9%	77.5	-8.8%	39.9	+76.5%

without constrastive loss resulted in a lower Chamfer distance and a higher F-score. One explanation for this result is that the contrastive loss may only be beneficial if the images were obtained from different camera perspectives, as was the case in Magistri et al. (2022). In our experiment, the potato tubers were photographed from the same camera perspective, which resulted in relatively similar 3D shapes, making it more difficult for the contrastive loss function to separate the embedding space. Two other observations from the ablation study that are useful for implementing CoRe++ on an operational harvester are: an instance segmentation algorithm is preferred over an object detection algorithm (as highlighted by the second ablation), and a lighter encoder network may be more favorable for high-throughput shape completion, but it comes at the expense of the accuracy (as highlighted by the third ablation).

4. Discussion

Of the six latent sizes tested, we observed that a latent size of 32 outperformed the other sizes in terms of 3D shape completion and volumetric estimate. This suggests that a moderate latent size strikes a balance between representational capacity and model complexity, allowing RGB-D images to be encoded more effectively while not overfitting. The latter was clearly demonstrated by the latent space interpolation, which generated the most realistic potato shapes and sizes for latent size 32. Compared to the literature, our observations are similar to those of Ahmed and Longo (2022), who found that a latent size of 28 was best for optimizing a convolutional variational autoencoder on spectral topographic maps. Since Ahmed and Longo (2022) did their research on 25 latent sizes, this suggests that even better results could have been obtained if we had tested more latent sizes. Future work could investigate such a finer-grained latent size analysis or further explore the relationship between neural network architecture and latent size. Such an analysis could potentially provide new insights into how to further optimize the latent space for 3D shape completion tasks.

With CoRe++ better 3D shape completion results were obtained compared to the linear regression model and the original CoRe implementation of Magistri et al. (2022). When we compare our results with those of Magistri et al. (2022), we can conclude that our results on potato are similar (in comparison with strawberry), or better (in comparison with sweet pepper). An important remark is that the obtained results may depend on the average complexity of the shape that has to be completed. Potato has on average a less complex shape than sweet pepper, which makes it easier for the network to learn a generic shape thus enabling a better performance. In future research, we want to test our CoRe++ model on crops or fruits with a more complex shape, such as pineapple, dragon fruit and Romanesco broccoli. We encourage fellow researchers to test our publicly available software on other 3D shapes within the agricultural domain or beyond.

Our research has provided valuable insights and steps towards the practical application of CoRe++ on a potato harvester. Nevertheless, we think there are potential improvements in software and hardware that could further improve the performance. Regarding software, we would like to explore the use of other 3D shape completion networks, such as the one by Magistri et al. (2024), who used a transformer network. The obtained results with this new transformer network on sweet pepper and strawberry were significantly better than those of the original CoRe implementation (Magistri et al., 2022). Another advantage of using a transformer network is that it is end-to-end trainable, which is advantageous over our CoRe++ network that needs to be trained in two stages. A hardware improvement that could potentially improve the 3D shape completion is equipping the conveyor with rotating rollers in the area where the camera is placed. The rotating rollers cause the potato tubers to rotate gradually as they pass underneath the camera, meaning that almost the entire shape of the potato can be photographed. This can both simplify and improve the 3D shape completion. In a scenario like this, it would also be useful to test whether the contrastive loss function has an improving effect on completing the 3D shapes of the potato tubers when they rotate.

To further test the applicability of CoRe++ on a potato harvester, it is important to conduct a more in-depth evaluation on potato tubers of different cultivars. This evaluation will provide a better understanding of the overall generalization performance of CoRe++ at the farm level. In such an analysis, it is also important to conduct the test on potato tubers without the colored thumbtack, as the thumbtack's color and shape can potentially affect the 3D shape completion result. A future evaluation could also benefit from performing the analysis on a more balanced dataset. In our study, potato tubers were randomly selected from the conveyor belt, but this led to an under-representation of the small and large tubers and those with irregular shape. In data-driven approaches such as ours, an unbalanced dataset would usually result in worse results on these minority cases, which is undesirable for a full-field harvest monitoring. Future research should also focus on reducing the degree of human subjectivity in selecting tubers on the conveyor belt by analyzing a batch of potatoes of different sizes at one time rather than manually selecting individual tubers. A final remark is that the performance of CoRe++ on a potato harvester will be affected by the instance segmentation algorithm that provides the necessary binary mask for CoRe++ to complete the 3D shape. Therefore, it is important to also investigate the sensitivity of CoRe++ to the segmentation outputs of the proposed instance segmentation algorithm.

5. Conclusions

In this study, we investigated a high-throughput 3D shape completion network for its applicability for estimating the volume of potato tubers on an operational harvester. Our research revealed that latent size 32 had the best 3D

shape completion result. With that latent size, our CoRe++ network had an RMSE of 22.6 ml on the volumetric estimate, and this was better than the RMSE of the linear regression (31.1 ml) and the original CoRe network (36.9 ml). We also found that the RMSE of CoRe++ could be further reduced to 18.2 ml when performing the 3D shape completion in the center of the RGB-D image. With an average analysis time of 10 milliseconds per potato tuber, CoRe++ enables a high-throughput 3D shape completion up to 100 potato tubers per second. Hence, we can conclude that our network is able to quickly and accurately estimate the volume of fast-moving potato tubers on an operational harvester. Future improvements lie in expanding the dataset, testing on more potato cultivars, implementing a transformer-based network, and implementing rotating rollers on the harvester's conveyor belt. We encourage other researchers to reuse our dataset and software.

Acknowledgements

We would like to thank Okada Farm for providing the potato field on which we acquired the RGB-D images. We thank Ting Jiang, Sylvain Grison, Yuto Imachi and Irena Drofová for their help with the image acquisition and ground truth measurements.

CRediT authorship contribution statement

Pieter M. Blok: Conceptualization, Methodology, Software, Data curation, Writing - original draft. **Federico Magistri:** Methodology, Software, Writing - review & editing. **Cyrill Stachniss:** Methodology, Software, Writing - review & editing. **Haozhou Wang:** Data curation, Resources, Software, Writing - review & editing. **James Burridge:** Conceptualization, Methodology, Data curation, Supervision, Writing - review & editing. **Wei Guo:** Conceptualization, Methodology, Funding acquisition, Project administration, Supervision, Writing - review & editing.

References

Ahmed, T., Longo, L., 2022. Examining the size of the latent space of convolutional variational autoencoders trained with spectral topographic maps of eeg frequency bands. IEEE Access 10, 107575–107586. doi:10.1109/ACCESS.2022.3212777.

Blok, P.M., van Henten, E.J., van Evert, F.K., Kootstra, G., 2021. Image-based size estimation of broccoli heads under varying degrees of occlusion. Biosystems Engineering 208, 213–233. doi:10.1016/j.biosystemseng.2021.06.001.

Cai, Z., Jin, C., Xu, J., Yang, T., 2020. Measurement of potato volume with laser triangulation and three-dimensional reconstruction. IEEE Access 8, 176565–176574. doi:10.1109/ACCESS.2020.3027154.

Chen, H., Liu, S., Wang, C., Wang, C., Gong, K., Li, Y., Lan, Y., 2023. Point cloud completion of plant leaves under occlusion conditions based on deep learning. Plant Phenomics 5. doi:10.34133/plantphenomics.0117.

Cheng, H.K., Chung, J., Tai, Y.W., Tang, C.K., 2020. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. arXiv:2005.02551.

David, E., Madec, S., Sadeghi-Tehran, P., Aasen, H., Zheng, B., Liu, S., Kirchgessner, N., Ishikawa, G., Nagasawa, K., Badhon, M.A., Pozniak, C., de Solan, B., Hund, A., Chapman, S.C., Baret, F., Stavness, I., Guo, W., 2020. Global wheat head detection (gwhd) dataset: A large and diverse dataset of high-resolution rgb-labelled images to develop and benchmark wheat head detection methods. Plant Phenomics 2020. doi:10.34133/2020/3521852.

Dolata, P., Wróblewski, P., Mrzygłód, M., Reiner, J., 2021. Instance segmentation of root crops and simulation-based learning to estimate their physical dimensions for on-line machine vision yield monitoring. Computers and Electronics in Agriculture 190, 106451. doi:10.1016/j.compag.2021.106451.

ElMasry, G., Cubero, S., Moltó, E., Blasco, J., 2012. In-line sorting of irregular potatoes by using automated computer-based machine vision system. Journal of Food Engineering 112, 60–68. doi:10.1016/j.jfoodeng.2012.03.027.

Fei, B., Yang, W., Chen, W.M., Li, Z., Li, Y., Ma, T., Hu, X., Ma, L., 2022. Comprehensive review of deep learning-based 3d point cloud completion processing and analysis. IEEE Transactions on Intelligent Transportation Systems 23, 22862–22883. doi:10.1109/tits.2022.3195555.

Ge, Y., Xiong, Y., From, P.J., 2020. Symmetry-based 3d shape completion for fruit localisation for harvesting robots. Biosystems Engineering 197, 188–202. doi:10.1016/j.biosystemseng.2020.07.003.

Geng, J., Xiao, L., He, X., Rao, X., 2019. Discrimination of clods and stones from potatoes using laser backscattering imaging technique. Computers and Electronics in Agriculture 160, 108–116. doi:10.1016/j.compag.2019.03.014.

Gómez, D., Salvador, P., Sanz, J., Casanova, J.L., 2019. Potato yield prediction using machine learning techniques and sentinel 2 data. Remote Sensing 11. doi:10.3390/rs11151745.

Hofstee, J., Molema, G., 2003. Volume estimation of potatoes partly covered with dirt tare, in: Proc. of the ASAE Annual Meeting 2003.

Huynh, T.T.M., TonThat, L., Dao, S.V.T., 2022. A vision-based method to estimate volume and mass of fruit/vegetable: Case study of sweet potato. International Journal of Food Properties 25, 717–732. doi:10.1080/10942912.2022.2057528.

Jang, S.H., Moon, S.P., Kim, Y.J., Lee, S.H., 2023. Development of potato mass estimation system based on deep learning. Applied Sciences 13. doi:10.3390/app13042614.

Kabir, M., Myat Swe, K., Kim, Y.J., Chung, S., Jeong, D.U., Lee, S.H., 2018. Sensor comparison for yield monitoring systems of small-sized potato harvesters, in: Proc. of the 14th International conference on precision agriculture.

Kurek, J., Niedbała, G., Wojciechowski, T., Świderski, B., Antoniuk, I., Piekutowska, M., Kruk, M., Bobran, K., 2023. Prediction of potato (solanum tuberosum l.) yield based on machine learning methods. Agriculture 13. doi:10.3390/agriculture13122259.

Lee, Y.J., Kim, K.D., Lee, H.S., Shin, B.S., 2018. Vision-based potato detection and counting system for yield monitoring. Journal of Biosystems Engineering 43, 103–109. doi:10.5307/JBE.2018.43.2.103.

Lee, Y.J., Shin, B.S., 2020. Development of potato yield monitoring system using machine vision. Journal of Biosystems Engineering 45, 282–290. doi:10.1007/s42853-020-00069-4.

Li, B., Xu, X., Zhang, L., Han, J., Bian, C., Li, G., Liu, J., Jin, L., 2020. Above-ground biomass estimation and yield prediction in potato by using uav-based rgb and hyperspectral imaging. ISPRS Journal of Photogrammetry and Remote Sensing 162, 161–172. doi:10.1016/j.isprsjprs.2020.02.013.

Li, D., Miao, Y., Gupta, S.K., Rosen, C.J., Yuan, F., Wang, C., Wang, L., Huang, Y., 2021. Improving potato yield prediction by combining cultivar information and uav remote sensing data using machine learning. Remote Sensing 13. doi:10.3390/rs13163322.

Long, Y., Wang, Y., Zhai, Z., Wu, L., Li, M., Sun, H., Su, Q., 2018. Potato volume measurement based on rgb-d camera, in: Proc. of the 6th IFAC Conference on Bio-Robotics. doi:10.1016/j.ifacol.2018.08.157.

Loop, C., 1987. Smooth subdivision surfaces based on triangles. Master's thesis. University of Utah, Department of Mathematics.

Lorensen, W.E., Cline, H.E., 1987. Marching cubes: A high resolution 3d surface construction algorithm, in: Proc. of the 14th Annual Conference on Computer Graphics and Interactive Techniques. doi:10.1145/37401.37422.

Magistri, F., Marcuzzi, R., Marks, E., Sodano, M., Behley, J., Stachniss, C., 2024. Efficient and Accurate Transformer-Based 3D Shape Completion and Reconstruction of Fruits for Agricultural Robots, in: Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA).

Magistri, F., Marks, E., Nagulavancha, S., Vizzo, I., Läebe, T., Behley, J., Halstead, M., McCool, C., Stachniss, C., 2022. Contrastive 3d shape completion and reconstruction for agricultural robots using rgb-d frames. IEEE Robotics and Automation Letters 7, 10120–10127. doi:10.1109/LRA.2022.3193239.

Marangoz, S., Zaenker, T., Menon, R., Bennewitz, M., 2022. Fruit mapping with shape completion for autonomous crop monitoring, in: Proc. of the IEEE International Conference on Automation Science and Engineering (CASE). doi:10.1109/CASE49997.2022.9926466.

Noordam, J.C., Otten, G.W., Timmermans, T.J., van Zwol, B.H., 2000. High-speed potato grading and quality inspection based on a color vision system, in: Proc. of the Machine Vision Applications in Industrial Inspection VIII.

Pan, Y., Magistri, F., Läbe, T., Marks, E., Smitt, C., McCool, C., Behley, J., Stachniss, C., 2023. Panoptic mapping with fruit completion and pose estimation for horticultural robots, in: Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). doi:10.1109/IROS55552.2023.10342067.

Pandey, N., Kumar, S., Pandey, R., 2019. Grading and defect detection in potatoes using deep learning, in: Verma, S., Tomar, R.S., Chaurasia, B.K., Singh, V., Abawajy, J. (Eds.), Communication, Networks and Computing, Springer Singapore. pp. 329–339.

Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S., 2019. Deepsdf: Learning continuous signed distance functions for shape representation, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Razmjooy, N., Mousavi, B.S., Soleymani, F., 2012. A real-time mathematical computer method for potato inspection using machine vision. Computers & Mathematics with Applications 63, 268–279. doi:10.1016/j.camwa.2011.11.019.

Salvador, P., Gómez, D., Sanz, J., Casanova, J.L., 2020. Estimation of potato yield using satellite data at a municipal level: A machine learning approach. ISPRS International Journal of Geo-Information 9. doi:10.3390/ijgi9060343.

Si, Y., Sankaran, S., Knowles, N.R., Pavek, M.J., 2018. Image-based automated potato tuber shape evaluation. Journal of Food Measurement and Characterization 12, 702–709. doi:10.1007/s11694-017-9683-2.

Su, Q., Kondo, N., Li, M., Sun, H., Al Riza, D.F., Habaragamuwa, H., 2018. Potato quality grading based on machine vision and 3d shape analysis. Computers and Electronics in Agriculture 152, 261–268. doi:10.1016/j.compag.2018.07.012.

Sun, C., Feng, L., Zhang, Z., Ma, Y., Crosby, T., Naber, M., Wang, Y., 2020. Prediction of end-of-season tuber yield and tuber set in potatoes using in-season uav-based hyperspectral imagery and machine learning. Sensors 20. doi:10.3390/s20185293.

Tomar, V.S., 2022. A critical evaluation of activation functions for autoencoder neural networks. Master's thesis. National College of Ireland, Dublin.

Xu, D., Chen, G., Jing, W., 2023. A single-tree point cloud completion approach of feature fusion for agricultural robots. Electronics 12. doi:10.3390/electronics12061296.

Zamani, D.M., Ghoşamiparashkohi, M., Faghavi, A., Ghezavati, J., 2014. Design, implementation and evaluation of potato yield monitoring system. International Journal of Technical Research and Applications 2, 36–39.

Zhang, H., Xu, F., Wu, Y., Hu, H., Dai, X., 2017. Progress of potato staple food research and industry development in china. Journal of Integrative Agriculture 16, 2924–2932. doi:10.1016/S2095-3119(17)61736-2.