Computers and Electronics in Agriculture 237 (2025) 110470

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

DeepCanola: Phenotyping brassica pods using semi-synthetic data and active learning

Larissa J.J. van Vliet [a,1], Kieran Atkins [a,1], Smita Kurup [c], Laura Siles [c], Jo Hepworth [d,e], Fiona M.K. Corke [a], John H. Doonan [a,*], Chuan Lu [b,*]

[a] National Plant Phenomics Centre, IBERS, Aberystwyth University, Aberystwyth SY23 3EE, UK
[b] Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, UK
[c] Plant Sciences and the Bioeconomy, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK
[d] John Innes Centre, Norwich Research Park, Colney Lane Norwich NR4 7UH, UK
[e] Department of Biosciences, Durham University, Stockton Road DH1 3LE, UK

ARTICLE INFO

Dataset link: https://github.com/kieranatkins/deepcanola, 10.5281/zenodo.14235543

Keywords:
Deep learning
Plant phenotyping
Semi-synthetic data
Active learning
Human-in-the-loop
Pod length

ABSTRACT

Phenotyping, the measurement of attributes or traits, is crucial in selecting superior cultivars for specific environmental situations. This is a time-consuming process when applied to large populations but can be accelerated through the use of deep learning, resulting in an algorithm that can phenotype images of specimens in negligible amounts of time. The primary issue with deep learning is the large quantities of high-quality training data required to make a viable phenotyping pipeline. To address this, we present a semi-synthetic training data generation system which significantly reduces the amount of human effort spent on data collection. We use active learning alongside this system to create *DeepCanola*, an instance segmentation model that successfully segments and measures the valves from *Brassica napus* pods. We demonstrate that the model accurately estimates the effect of different winter cold treatments on a range of different cultivars and crop types as effectively as manually curated measurements. Furthermore, the resulting model is effective on data from various experimental settings and on different, but related, species such as *Arabidopsis thaliana*, *Allaria petiolate* (garlic mustard) and *Raphanus raphanistrum subsp. sativus* (radish). This robust tool could be easily scaled, thereby accelerating breeding or fundamental research programs. Code and model weights: https://github.com/kieranatkins/deepcanola.

1. Introduction

Phenotyping, the accurate measurement or estimation of traits from individuals, is a very powerful technique to explore and quantify differences within diversity populations, which display genetic variation and form the foundation of many breeding programs. In a crop breeding context, traits of particular interest often include the size and number of fruits as these are often the economically relevant part of the plant. In ecological terms, fruits reflect the effort that a mother plant put into reproduction. In both cases, their production and quality are affected by the plant's tolerance to environmental stresses, such as the ability to cope with warming climates and resistance against naturally occurring diseases. The ability to accurately, quickly and reliably measure traits such as fruit size is then, clearly, of high importance.

Brassica napus has given rise to several major crops, including the oilseed canola/rapeseed, forage rape, swede/rutabaga and industrial oilseeds. Breeding cultivars suitable for different environments has led to locally adapted and highly productive commercial crops (Diepenbrock, 2000). Phenotypic traits commonly measured in these crops are flowering time (Calderwood et al., 2021; Williams et al., 2023) and seed yield (Siles et al., 2021). Flowering time (when the first flowers open) is easily scored using traditional manual assessment and the genetic basis has been dissected in detail (Schiessl et al., 2017; Calderwood et al., 2021). Yield, measured at harvest time, is a complex composite trait that is influenced by the environment and by multiple genes acting at various stages throughout development. Components of yield include pod number and the amount of seed per pod.

B. napus pods are botanically defined as siliques, derived from the two ovary carpels. Each pod is formed by two walls, also known as valves, which contain the seeds, a pedicel that attaches the pod to the plant stem, and a beak at the distal end. The beak is a seedless

[*] Corresponding authors.
E-mail addresses: jhd2@aber.ac.uk (J.H. Doonan), cul@aber.ac.uk (C. Lu).
[1] Equal contribution.

https://doi.org/10.1016/j.compag.2025.110470
Received 5 April 2024; Received in revised form 11 April 2025; Accepted 25 April 2025
Available online 11 June 2025

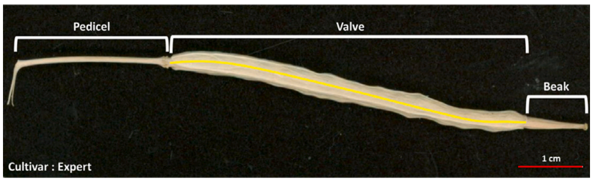


Fig. 1. An example Brassica pod from the cultivar expert with the pedicel, valve and beak regions annotated. The length of the valve region is shown by the yellow line.

structure derived from the gynoecia style (Hossain et al., 2012; Gulden et al., 2008) and has different shapes and sizes depending on the genotype and the environment experienced during the flowering season. Variation in valve size, and the ultimate output of the fruit, does not necessarily correlate with the beak size. Due to the differences observed in different genotypes and the fact that this pod structure does not contain seeds, it is more accurate to measure valve length rather than the whole pod length to relate it with seed content/production (Siles et al., 2021). An example pod with annotated valve and beak regions is illustrated in Fig. 1.

For image-based size analysis, pods are usually removed from the mother plant, arranged as isolated pods and imaged against a high-contrast background. This arrangement facilitates the use of classical computer vision tools as algorithms such as *connected components* can isolate disconnected segmented objects for analysis after thresholding. However, this approach typically requires that each pod is arranged to have no occlusion (overlap) or contact of any kind. Meticulous arrangement of material increases the amount of human effort and time needed during the imaging phase, and therefore reduces such a system's utility in a high-throughput phenotyping context. Furthermore, classical CV can only reliably phenotype the pod as a whole, as opposed to the yield-relevant valve, as there are few reliable landmarks that can be used to robustly segment the valve region using classical CV. Therefore, classical CV algorithms have limitations for high-throughput valve phenotyping. While classical CV struggles with these challenges, deep learning offers a versatile and robust approach to the task.

Deep learning presents a paradigm shift compared to classical CV pipelines, as models are trained to replicate the relationship between inputted images and hand-collected training data, to generate new outputs on unseen images. Deep learning models have been used previously for a variety of fruit phenotyping tasks. For example, Lu et al. (2022) developed a system where a CNN object detector (YOLOv3) identifies and counts leaves and pods in images of soybean plants. A generalised regression neural network (GRNN) then uses this information to accurately estimate seed yield. This approach offers advantages such as in-situ estimation, reduced researcher bias, and increased phenotyping throughput.

Although object number is an informative trait, there is rich information, such as size, shape, colour and location, about each object that these detection/counting methods do not necessarily capture. Instance segmentation offers three key capabilities for extracting object traits: object detection, classification, and pixel-level segmentation. A proven and robust model choice for these tasks is Mask R-CNN, a model built on top of an existing powerful object detector (Faster R-CNN) with a branch for mask prediction. Mask R-CNN utilises a CNN backbone to extract image features (He et al., 2017). These features are then used by: 1) a region proposal network (RPN) to identify potential regions of interest, 2) an object detection branch to infer a bounding box and class for each region proposed by the RPN (Ren et al., 2015), and 3) a mask prediction branch to generate a detailed pixel-level mask for each region of interest, outlining its shape.

Once object masks have been extracted, extracting phenotypic data such as area (a 2D substitute for biomass) and length is comparatively computationally inexpensive. Therefore, Mask R-CNN shows considerable promise for extracting phenotypic data from different crops. Su et al. (2020) trained two Mask R-CNNs to identify individual wheat spikes and Fusarium head blight, an infection that affects wheat grains. The first model detected individual wheat spikes with a spike classification accuracy of 77.76%, whereas the second model was able to recognise the diseased grains within the spike with an accuracy of 98.81%, with a mask average precision (AP) of 0.57. Eventually, the area of the spike and the area of the disease were used to estimate a disease severity score. The lower accuracy of the first model is partially caused by occlusion of wheat spikes. Due to the nature of field images, overlapping wheat spikes are unavoidable. This overlapping can be difficult to resolve using Mask R-CNN alone, especially if the model was trained on a dataset with limited examples of such overlap.

Occlusion of one object by another is a common challenge for accurate object detection. To address this, Liu et al. combined Mask R-CNN with a DBSCAN clustering algorithm to resolve overlaps between leaves in seedlings. While Mask R-CNN alone reached a detection AP of 0.877 (with a 0.7 confidence threshold in object classification), adding DBSCAN reduced false positives and improved the detection AP to 0.892.

An instance segmentation model's output is much more fine-grained and can generate more precise information than their object detection-only counterparts, allowing for the generation of more informative data. This comes at a cost, however, as the training step requires manually annotated masks rather than simple classifications or bounding boxes. Hand-collection of mask data is a labour-intensive and time-consuming process. To reduce the annotation burden, synthetic data with masks is often generated using (1) semi-synthetic approaches, which compose real objects with augmentations into new scenes, and (2) fully synthetic methods, which create data via computational models.

For example, Toda et al. (2020) used semi-synthetic data to train a Mask R-CNN which was then evaluated on real images. By annotating a relatively small number of barley seeds and using them to create semi-synthetic images, they were able to reach an average recall and average precision of 0.96 and 0.95 in object detection, respectively, and a mask AP of 0.59 on real images. The approach was transferable to other crops, such as rice, lettuce, oats and wheat, indicating wider applicability. A similar approach used to segment individual soybeans and extract phenotypic traits for high-throughput data extraction achieved a high accuracy in object detection. However, it struggled to recognise burst pods (Yang et al., 2021).

Unlike composition-based semi-synthetic data generation, recent fully synthetic methods often rely on 3D modelling tools like Blender to create detailed 3D scenes for generating synthetic images. For example, Napier et al. (2023) utilised L-systems, a grammatical framework encoding plant morphological development, to generate 3D models of wheat heads, which were then rendered into synthetic in-field images for training. Lately, denoising diffusion models, such as Stable Diffusion, have demonstrated their effectiveness in style and domain transfer from given images, surpassing earlier GAN-based methods for image generation (Wu et al., 2023a). These models rely on prompts, often carefully crafted text, to describe domain-specific features for style adaptation. har (2024) applied Stable Diffusion and LoRA to L-system based 3D models of Arabidopsis rosettes, producing realistic datasets for training Mask R-CNN in leaf detection and segmentation.

Fully synthetic image generation, such as diffusion-based models, are capable of producing convincing outputs in negligible time. This could be invaluable in the crop phenotyping domain where image collection can protract over seasons (or even years in the case of perennials) and many objects may need annotation within a single data sample. However, generative AI models are often significantly computationally expensive to train, requiring large datasets and substantial resources. Furthermore, generative AI models can be susceptible to hallucination, where inaccurate or imprecise outputs are confidently produced. Generating training samples from 3D models, however, does not typically have the same issues as generative AI, and can be used to create high-quality training data, without large training datasets, nor being susceptible to hallucination. Instead, 3D models require detailed modelling of plant structure. The specific task discussed in this study, the imaging of Brassica pods (and parts of pods) isolated from the mother plant on a consistent largely uncluttered background, is a more constrained task than those to which generative AI and 3D modelling are typically applied. Therefore, these complex modelling systems would likely result in unnecessary computational overhead to produce outputs that could be similarly produced using classical CV algorithms. For a more appropriate training data generation system for the target domain, we chose to annotate the valve regions within images of ordered and isolated Brassica pods, and programmatically rearrange the annotated pods with randomised positions and orientations, or diverse combinations or crowding of pods across different images making a much larger set of high-fidelity semi-synthetic samples for training.

Although it may not be as complex or produce images with the same fidelity as synthetic data generation using large or complex models in some domains, our proposed method offers distinct advantages over fully synthetic data generation. Firstly, the computational overhead is reduced, requiring only pod images with annotations to be collected, along with the development of a simple pipeline to place these images within novel scenes of the researcher's choosing. Utilising true images ensures the creation of novel training samples based on real data as collected by crop scientists and is incapable of producing hallucinated outputs.

Furthermore, additional training data could be generated by geometric transformation and non-linear shape alterations, akin to the work by Thompson (1961). D'Arcy Thompson described how variation between parts of organisms (i.e. organs) of related species can be described by relatively simple geometric transformations. We reasoned that introducing simple geometric transformations into semi-synthetic sample generation could therefore expand the model's generalisation potential, allowing the model to generalise to related cultivars, and wider relatives in the Brassicaceae family. To iteratively understand the limitations of a model trained using this system and to expand its capabilities, we therefore employed such a semi-synthetic dataset alongside an active learning system.

Active learning is a time-efficient computational method to retrain models, iteratively improving its ability to identify atypical features or to extend the range of a model to recognise related features. Active learning typically involves selectively choosing which additional data will deliver the largest gain in accuracy with the least amount of labelled data (Blok et al., 2022; Granland et al., 2022). This can reduce the amount of manual annotation required. Active learning systems have successfully been used in the phenotyping of cereal spikes and stalks, including wheat and sorghum (Kumar et al., 2019; Chandra et al., 2020). However, these studies focused on object detection only. Rawat et al. (2022) used different strategies of uncertainty-based active learning for semantic segmentation of apples, wheat and rice and showed that it had little benefit compared to random sampling. This reflects the significance of task-specific data augmentation to address real-world challenges like occlusion. Whilst strategies for active learning perform differently on different tasks, the theory of using the model's output to identify which data is best suited for improving accuracy is important for diverse datasets such as those from plants.

To reduce the cost of manual labelling and address the limitations of typical automatic uncertainty measures in active learning, we employed a hybrid approach combining synthetic data augmentation with a human-in-the-loop (HITL) strategy. Human experts play a vital role in selecting validation dataset (conscious of biases) and qualitatively assessing model performance on predicted instance segmentation masks. Their insights directly inform the generation of synthetic images, tailored to the specific challenges of our instance segmentation tasks, ensuring effective semi-synthetic data generation.

In this study, we developed a robust method to accurately identify and measure pod valves from *B. napus* in both organised and disorganised scenes, based on a deep learning instance segmentation model and a semi-synthetic training dataset. *B. napus* and its relatives display a range of pod morphologies that vary in terms of both size and shape (Łangowski et al., 2016), providing a good test for the model's ability to generalise across different types of pod. To train the model, images of the pods of *B. napus* were manually annotated and extracted prior to creation of a semi-synthetic image dataset. The semi-synthetic image dataset was used to train a Mask R-CNN model, which was subsequently validated using a set of real images to evaluate the model's performance. We use uncertainty sampling, an active learning strategy where new samples are included based on low-confidence results from previous iterations, to allow for the model to generalise better on more diverse samples. By using this paradigm, where new training data was selected based on the model's mistakes, new semi-synthetic datasets were created for further training. After 4 rounds of active learning, the model outputs were judged to be satisfactory, the predictions were used to estimate the length of the valve and tested against manually measured data to check for the accuracy of the model and consistency with the manually obtained experiment results. We then validated our final model, which we term DeepCanola, on images from diverse experiments, imaging setups, and species to assess its versatility in applying learned patterns to novel biological systems and its adaptability across different imaging scenarios.

2. Materials and methods

2.1. Experimental design and sampling

The initial images were taken from experiments designed to analyse the trait variation associated with the transition to flowering and subsequent developmental events such as raceme elongation in a range of cultivars (Williams et al., 2023). These experiments are named BR9, BR11 and BR17. For BR9 and BR11, seedlings were vernalised and then grown to maturity in a mechanised greenhouse or Smarthouse. BR17 was designed to understand the developmental response to defined periods of cold as a seedling. Seedlings from each cultivar were subjected to a cold period of either 5 °C or 10 °C before transfer to a common growing environment within a Smarthouse where they were allowed to flower and set seed (Williams et al., 2023). These and other sources of material are summarised in Table 1. The pod sampling strategy was based on (Siles et al., 2021) and approximately 20 typical mature pods were harvested from the primary stem of each individual.

2.2. Image acquisition

Isolated pods from each individual plant were imaged. An example pod is shown in Fig. 1. Pods were either neatly ordered into rows (where pods are all pointing in the same direction and separated to eliminate occlusion) or disordered (where pods lay across one another to variable degrees).

To test the final model's generalisation potential on more complex images, we took advantage of image data that had been collected during the systematic disassembly of mature plants from experiment BR17. This involved the removal and scanning or photographing of each branch with its pods still attached. Additionally, images were obtained from other Brassicaceae, including *Raphanus raphanistrum* subsp. *sativus* (radish) and *Alliaria petiolate* (Garlic mustard), which were collected from farmland (Norfolk) and natural wooded areas (Aberystwyth) respectively. To test whether the approach could be extended to data collected for other purposes, (i.e as could be obtained by the general public) images were collected from the iNaturalist community (iNaturalist community, 2023) for pods of Arabidopsis, which produce fruits with some similarities to Brassica pods.

Table 1
Summary of data sources used in this study, showing number of images, annotations and related details. (a) Real-world datasets include datasets used to create the synthetic training data and datasets used for model validation. (b) Generated datasets include the semi-synthetic data created at each step of the active learning process. # Images contains the number of images in the full dataset. In (a) # Annot. contains numbers of valve mask annotations collected from the dataset, in (b) # Annot. contains number of annotated pod valves present in full generated dataset.

Dataset	# Images	# Annot.	Image details	Biological details
BR9	211 ordered 97 disordered	673	Ordered and disordered pods, 2552 × 3508 pixels, imaged on flatbed scanner	Randomised, triplicated design including 89 lines of the OREGIN B. napus diversity fixed foundation set comprising winter, spring, Chinese, kale and swede; 5 °C vernalisation
BR17	412 ordered	332	Ordered pods, 2551 × 4200 pixels, imaged on flatbed scanner	Randomised, triplicated design including 73 lines of the RIPR B. napus diversity fixed foundation set, either 5 °C or 10 °C vernalisation, collected with manual length measurements
BR11	35 ordered 36 disordered	0	As BR9	As BR9
Misc.		0	Images from various locations and sources	Various Brassica relatives

(a) Real-world datasets

Dataset	# Images	# Annot.	Generation details	Data source
Dataset 1	100	4410 (30–60 per image)	2552 × 3508 pixels, pods randomly placed on background	BR9
Dataset 2	100	4501 (30–60 per image)	2552 × 3508 pixels, pods randomly placed on background with random blur	BR17
Dataset 3	200	8356 (20–60 per image)	2552 × 3508 pixels, pods randomly placed on background with random blur	BR9 & BR17
Dataset 4	1000	44823 (average 45 per image)	2552 × 3508 pixels, pods randomly placed on background with synthetic blur representative of real-world images, higher quantity of objects and images, stronger object and background augmentation	BR9 & BR17

(b) Generated datasets

2.3. Model development

Development of strong deep learning models requires a large amount of high-quality training data. To accumulate sufficient training data, we employed a semi-synthetic training data generation schema, which was iteratively expanded and improved using active learning. The generation of semi-synthetic training data allowed for (1) a reduction in the time required to collect training data and (2) tuning of diversity of pod types and morphologies with object-level augmentation to create an informative training dataset. Basing training on semi-synthetic samples required pods to be individually annotated separate from other plant biomass so that they could be arranged, at will, on a selected background. These collected pod image pools with associated annotations formed the basis of our semi-synthetic training data generation. The pools, along with the generation parameters, were expanded and adjusted, responding to the results of the active learning process.

The active learning process utilised uncertainty sampling to identify weaknesses in a given model's outputs. Uncertainty sampling is a commonly used active learning process whereby poorly performing samples, identified by their low confidence scores or incorrect segmentation, are identified and reintroduced in later rounds of training. We also utilised human crop-breeding experts who qualitatively assessed outputs, for example, identifying specific weakness in model performance or where the semi-synthetic training data was unrepresentative of real-world samples. This workflow is outlined in Fig. 2.

2.3.1. Semi-synthetic image generation

The semi-synthetic images were designed to capture and reflect the placement, angle and distribution of randomly placed pods on a surface, as would be the case for a large-scale breeding program. The process is as follows: Firstly, a random background is selected from the background pool and an empty mask of the background initialised. Then, a number of pods n to be placed in the image is randomly selected. Each pod is randomly selected from the pool of pod images, segmented from the high-contrast black background, and placed at a random position, orientation and scale with the bounds of the image. The associated valve masks are then stored with the same position, orientation and scaling applied to them. This is repeated until the required number of pods is reached. Once all are placed with the image, occluded pods have a level of depth-of-field blur applied to them to imitate real-world imaging. This workflow is illustrated in Fig. 2. The training sample is then stored, and the next sample is created until the desired dataset size is reached.

2.3.2. Iterative expansion of the semi-synthetic dataset for model training

Each round of the active learning process involved collecting new data for the model to learn, resulting in four datasets. In order to create each semi-synthetic dataset, a pool of backgrounds and pods with annotations was needed. In all rounds, pod annotations were collected using Fiji (Schindelin et al., 2012).

To create the initial pod image pool, the valve regions within 20 images containing organised pods from seven different BR9 cultivars

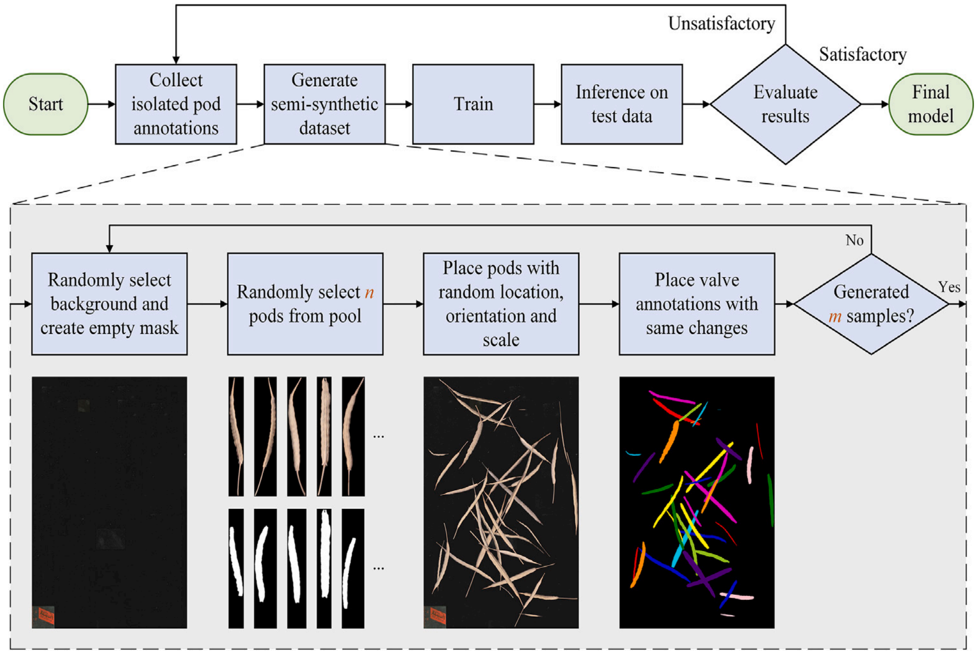

Start
Collect isolated pod annotations
Generate semi-synthetic dataset
Train
Inference on test data
Evaluate results
Unsatisfactory
Satisfactory
Final model
Randomly select background and create empty mask
Randomly select n pods from pool
Place pods with random location, orientation and scale
Place valve annotations with same changes
Generated m samples?
No
Yes
...
...

Fig. 2. Active learning and semi-synthetic generation workflow — Workflow of the active learning process from dataset generation to training to validation. Exploded sub-illustration of the data generation process is also shown, starting from the background, pod and valve annotation pools, resulting in a generated semi-synthetic training sample. Training sample shown is from Dataset 4 (4th iteration of the active learning process). Note: Colours shown in the example semi-synthetic annotations are illustrative, colours are reused and therefore do not represent a single instance.

were manually annotated, which took about 15 man-hours to complete. For the second round of training, another 20 images were annotated and for the third round of training the pods included in both the first and second rounds of training were combined. After non-singletons were removed, the pod pool consisted of 673 images of individual pod objects and their associated masks. To create an image pool of backgrounds, 15 patches that had no pods present were manually cropped from real images, after which the patches were randomly chosen, rotated and placed in an empty image, which resulted in four background images. 100 semi-synthetic images were generated for the first dataset, and used to train the first model, however, during expert evaluation overfitting was observed on the pod phenotypes present in the BR9 dataset, showing poor performance in more extreme pod shapes.

The second round of training was designed to address the model's performance issues on extreme pod shapes, therefore more diverse and larger Brassica pods images and annotations were collected from the BR17 dataset. Manual annotation took approximately 24 person-hours of work and resulted in 332 additional pod objects with valve masks. 100 semi-synthetic images were generated using the new pod images pool and used to train a model. During evaluation, overfitting on the newer pod types was observed, with poorer results on pod phenotypes from the first dataset. This demonstrated that samples need to be blended from previous rounds to combat the overfitting that occurs when only training on the new samples. Issues were also observed where the model returned true positive results for "noise" within the images (such as dusty fingerprints, debris and pieces of the ruler). To address this, another seven patches that contained real-world noise were cropped and randomly placed on duplicates of the first four background images, which resulted in a total background pool of eight images.

The third round combined the pools of the first two rounds in order to combat overfitting, while expanding the size of the generated dataset to 200 samples. The resulting model showed a notable increase in performance, however during evaluation there were still issues related to extreme pod shapes, unrealistic blur and difficulty to detect pods in dense scenes.

To address these issues, we create a fourth training round with a focus on capturing the pod morphological diversity by creating an additional synthetic dataset, with the following changes.

1. Five-fold increase in the generated image count, increasing from 200 in round 3, to 1000 images.
2. Independent random scaling in the x&y directions, allowing free-axis morphological stretching rather than fixed scaling. This augmentation mimics the work by Thompson (1961), and should allow the model to better generalise towards extreme pod shapes of different cultivars.
3. Background lightness augmentation to add background variation to make the model more robust against changes in background lightness. To achieve this, backgrounds had a random level of gamma adjustment before placement of pods began.
4. Improved blurring system designed to better mimic the depth-of-field blurring effect caused by pods being at different distances to the camera lens when overlapping. To achieve this, depth-of-field blur was iteratively added depending on the number of pods occluding each pod.

These changes improved the results on our generated test-sets and on the real-world test-sets.

2.3.3. Model training and evaluation

We used a Mask R-CNN instance segmentation model (He et al., 2017), with a ResNet-50 + FPN backbone. For each model, the initial learning rate was set to 0.005, the momentum to 0.9 and weight decay 0.0005. All models were trained for 12 epochs, with learning rate decay factor (gamma) of 0.1 at epochs 8 and 11. The first, third and fourth models were initialised on weights pre-trained on the MS COCO dataset (Lin et al., 2014). The second model was initialised on the weights of the first model. All images were rescaled to the shortest edge being 1914 pixels long, whilst keeping the original aspect ratio. For training, this resulted in images of size of 1914 × 2631 (downscaled from 2552 × 3508). This downscaling kept memory usage down. Training details are listed in Table A.1.

We chose a consistent set of hyperparameters in all model training in order to keep the comparison between the different training datasets as fixed as possible. Consistent hyperparameters allowed us to observe how changes in the semi-synthetic dataset generation process affected model performance. Therefore, the default hyperparameters for Mask R-CNN training, provided by torchvision, were selected for this reason. We also observed little change when attempting to fine-tune the initial learning rate. 12 epochs of model training was selected as all model loss and validation metrics had converged and stabilised by this time, without overfitting.

After every training cycle, the model's performance was evaluated using the validation set to calculate the average precision (AP) and average recall (AR). These metrics provide insights into the model's ability to detect pod valves accurately. AP measures the proportion of correctly identified valves among all detections made, calculated by the ratio of True Positives (correctly identified pods) to the total number of positive predictions (True Positives + False Positives). AR measures the proportion of correctly identified pods among all pods in the test samples, calculated by the ratio of True Positives to the total number of actual positive objects.

To provide a comprehensive evaluation, AP and AR were averaged across different Intersection-over-Union (IoU) thresholds. IoU measures the overlap between the predicted bounding box/mask and the ground truth. IoU thresholds denote the minimum IoU of a detected object compared to the ground truth before being considered a True Positive. Common IoU thresholds include 0.5, 0.75 and an average from 0.5 to 0.95, with increments of 0.05. We report the latter 0.5:0.95 average, as it reports an average of model results from a range of IoU quality thresholds, giving a more rounded view of model performance.

The active learning process requires both model assessment and training data selection throughout the training process. To assess the model's performance, the BR17 dataset was used. How well the model performed on this selection of images informed us as to which pods should be included in the generation of the semi-synthetic dataset. After the semi-synthetic dataset was created, the model was trained and then assessed again. This process continued until a model with a desirable level of performance was achieved. The model was assessed quantitatively by metrics previously outlined, and qualitatively by assessing the output masks to see where and for what pods there are issues like false negatives, where pods are completely ignored, false positives, where the model detects pods valves that do not exist, or poor quality masks that do not properly segment the pod.

All experiments were performed on a workstation with an Intel Core i7-11700K CPU, 32 GB RAM and a single NVIDIA GeForce RTX 3080 with 10 GB VRAM

2.4. *Phenotype data extraction*

2.4.1. *Manual measurement*

To assess the model's accuracy in measuring valve lengths, we compared its computed measurements with manual reference values obtained of the BR17 images. Manual measurements in Fiji (Schindelin et al., 2012) involve setting the scale using the ruler present in the image to establish a pixel-to-millimeter conversion factor. Then, a line is drawn along the valve's central axis, following its curvature from base to tip (excluding the beak, as shown in Fig. 1). This manual approach serves as the ground truth for our evaluation. This process takes about 15 min per 20 pods.

2.4.2. *Model prediction*

For each individual object identified the model returns a bounding box that surrounds the predicted object, a confidence score for classification, and a mask that is located within the bounding box specifying the pixels predicted to be part of the object. We selected a confidence score threshold of 0.75 to ensure the quality of results returned by the model. Detections with a lower confidence score than the selected threshold tended to be false positive detections, often containing two or more pods in a single detection resulting in an incorrect mask. Selecting a higher confidence score threshold resulted in the suppression of true positive results. Additionally, each pixel within the bounding box has a mask probability. Pixels with probabilities exceeding 0.5 were considered part of the object.

To calculate length, the binary output mask was skeletonised, a mathematical operation that effectively erodes the sides of a binary image (the mask, in this case) and returns a central line, equidistant from the edges, that best approximates the object's topology. In the case of pods, this gives a central line that follows the curvature of the pod and extends from one end of the valve region to the other (see Fig. 1). The Euclidean distance between each pixel along this line gives a final length value for the valve length. The area was then calculated by summing the number of pixels in the mask.

2.5. *Testing methodology*

To investigate the progression of the models through the active learning process we split each dataset into random subsets of size 80% for training and 20% for testing. This allowed us to calculate the mask and box AP and AR metrics on unseen test data. We then evaluated each iteration against all the test sets of the previous iteration (Table 1) to assess how well the model performs on similar semi-synthetic datasets. However, to understand real-world performance, we tested against three real-world datasets: BR17, BR9 ordered/disordered and a species diversity panel.

BR17 represents a large scale experiment of ordered Brassica pods with hand-collected length measurements. This dataset allows us to directly compare the measurements generated by the model against ground-truth hand-collected data. While giving a good evaluation of the model's potential, BR17 only evaluates the model's performance on ordered images.

Disorder is a crucial factor to consider in biological samples. In high-throughput phenotyping many samples are naturally disordered, however ordering requires extra human intervention, therefore a phenotyping system must be able to cope with disorder. Dataset BR9 captures the difference between the two concepts. BR9 has two main image types: *Ordered* pods arranged in a grid fashion with zero overlap (the initial source of our annotated pods, see row 1 of Fig. A.1); *Disordered* pods placed without ordering and high amounts of overlap (see row 2 of Fig. A.1). We perform the same evaluation for BR17 on BR9 to assess how disorder affects the R^2 correlation between manual and deep learning phenotype data. This comparison is of particular interest because we can evaluate how each model responds to disorder by using exactly the same pods but in different states of disorder/order.

To evaluate the model's broader applicability, we collected smaller diverse datasets, representing different experiments, image acquisition settings, even non-experimental images. The model then generated outputs for these datasets, and we qualitatively assessed the results.

3. Results

3.1. *Model refinement through active learning*

We analysed each model iteration against the test sets generated at every training round. This allowed us to check how the model's detection performance evolved throughout the active learning process. Table 2 reports the box and segmentation AP for valve detection at each iteration. For more details on dataset generation, see Section 2.3.2 and Table 1.

Table 2 outlines the models' improvement against each step's test set over the course of the learning process. The first two training iterations highlight the problem of overfitting when using only the image pool from new annotations; the model excels on the newly generated dataset but performs poorly on older ones. Our solution was to combine new

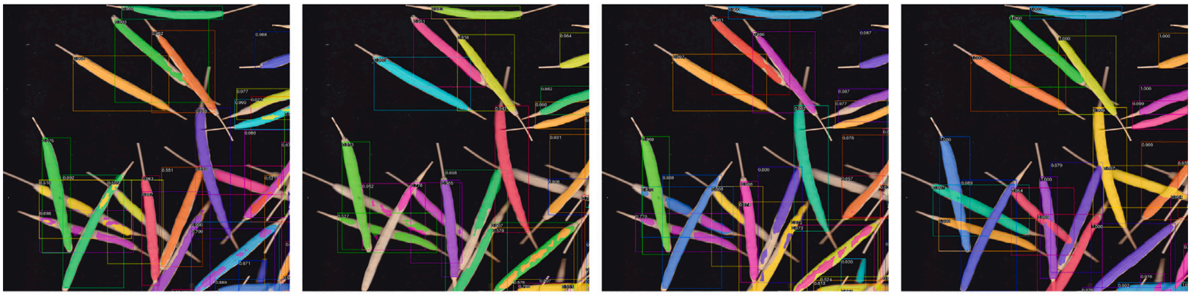
Table 2
Model performance (Detection/Segmentation AP) on each semi-synthetic dataset through the active learning process. Post-hoc analysis on earlier models are shown in grey. Best results shown in bold.

	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Model 1	0.618	0.304	0.528	0.406
Model 2	0.558	0.51	0.565	0.471
Model 3	0.661	0.485	0.615	0.511
Model 4 “DeepCanola”	**0.715**	**0.59**	**0.685**	**0.684**

(a) Detection/Bounding Box AP (0.5:0.95) results

	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Model 1	0.421	0.28	0.418	0.287
Model 2	0.403	**0.449**	0.456	0.349
Model 3	0.466	0.437	0.493	0.38
Model 4 “DeepCanola”	**0.483**	0.444	**0.515**	**0.583**

(b) Segmentation/Mask AP (0.5:0.95) results



(a) Model 1 (b) Model 2 (c) Model 3 (d) Model 4

Fig. 3. Active learning progression example — Example patch of test image from disordered BR9 dataset showing results from each of the steps of the active learning process. (a) Model 1 is the initial version of the model containing a limited pod image pool, without refined generation parameters. (b) Model 2 is built upon Model 1 but had overfit by being trained further on exclusively newly added samples. (c) Model 3 contains a larger dataset, with the combined image pools of Model 1 & 2 and achieves better results. (d) Model 4 (DeepCanola) builds upon Model 3 by expanding the size of the generated dataset five-fold and includes improved data generation parameters. Colours are for visualisation purposes only. Colours represent each instance of pod valve detected by the model, and is used for both the bounding box and segmentation. Note that colours are cycled through and reused for multiple instances. Extra examples are shown in Fig. A.2.

and old image pools when generating a semi-synthetic dataset throughout the training process. Despite the similarity between Datasets 1 and 2 (pods from related species), this mixing proved essential for high-quality results and generalisation, as demonstrated by the improved performance of Model 3. This is visualised in Fig. 3, we show a portion of an image from the BR9 dataset that visualises the model progression over the active learning process. Initially, we see good segmentation on isolated pods, but when pods are in close proximity or occluding one another then the mask generation fails to isolate one pod’s valve from another. Model 2 was designed with more diverse samples, but was trained only on samples generated from the new pods, and as such we see some improved segmentation in some pods, but a regression in others. Model 3 combined all pods and expanded the number of training samples, and as such we see improvements, but overall poor results in situations of overlap. The final model largely addresses these problems, resulting in a markedly better handling of segmentation in overlapping and neighbouring samples.

Model 4’s performance improvement stems primarily from the substantial increase in training images, achieved without additional human annotation effort. Our Thompson-inspired augmentation strategy generated a semi-synthetic training dataset with greater sample variation and better representation of real-world variation, boosting the model’s robustness against morphological variability. This is reflected in the increase in both detection and segmentation results, with detection AP rising from 0.615 to 0.685 and segmentation AP from 0.493 to 0.515 between Model 3 and Model 4. Based on these improvements, Model 4, henceforth referred to as DeepCanola, will be used for real-world data analysis.

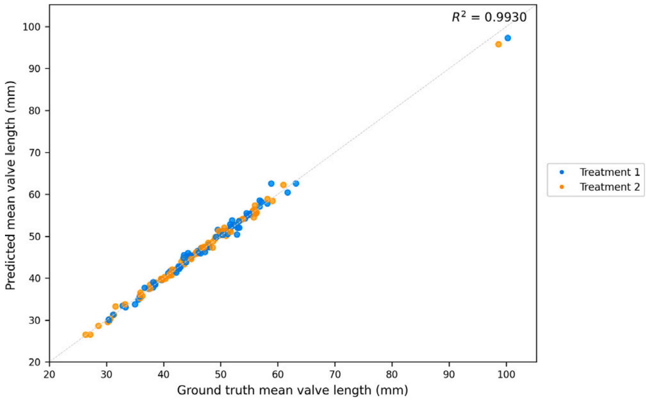


Fig. 4. Length prediction accuracy — Average valve lengths for each cultivar and treatment manual measurement (x axis) against average values predicted by DeepCanola (y axis), with treatment colour-coded. The target $y = x$ line is shown as a grey dotted line.

3.2. Results on ordered images

Dataset BR17 has a large quantity of hand-collected length data, and as such provides a strong test of a deep learning model trained on semi-synthetic data. Fig. 4 plots average valve lengths (per cultivar/treatment combination) obtained from manual measurements and those inferred by the model. With the $y = x$ line being a perfect model, this demonstrates the model’s ability to detect the variability in the pod-length phenotype, despite the difference in the size extremes being

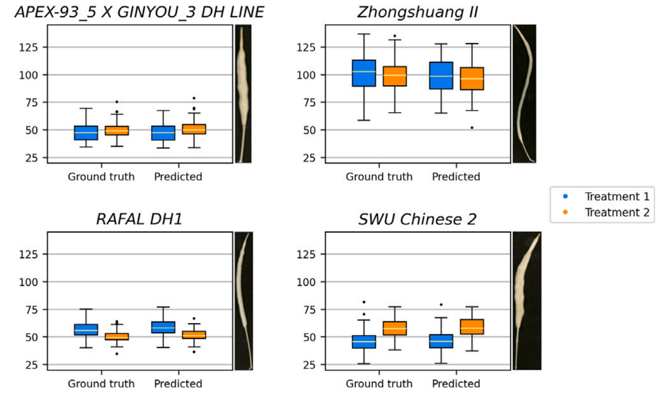


Fig. 5. Select cultivar ground-truth and predicted valve lengths — Boxplots comparing 5 °C (blue) and 10 °C (orange) vernalisation treatments on valve length in four cultivars (experiment BR17). Ground truth measurements were acquired manually. Example cultivars were selected to highlight morphological variation. See Fig. A.6 for the complete output for BR17.

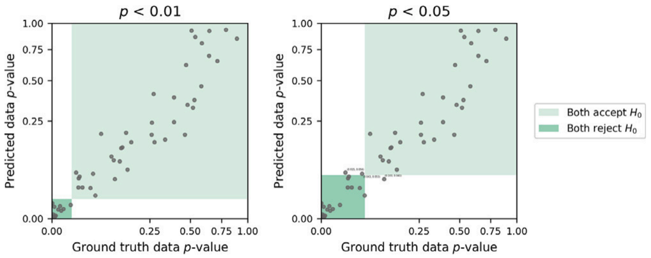


Fig. 6. Concordance of treatment effect significance on valve length — For each cultivar/treatment, related samples t-tests were conducted to assess the treatment effect on average valve length as determined by DeepCanola estimates and manual measurements. Scatter plots visualise the resulting p-values (square root transformed) for both methods. Agreement on the hypothesis test visualised in green, with agreement on rejection of the null hypothesis shaded in dark green, and agreement on acceptance of the null hypothesis shaded in light green. White regions highlight disagreement on hypothesis test between methods, with p-values reported.

almost 3-fold. On these ordered images the model achieves an R^2 = 0.9930 for the length calculations. This high value demonstrates the strength of our model and processing methods for extracting accurate phenotypic data from such image types.

To test DeepCanola's ability to measure the effect of environmental treatment on different cultivars, we compared predicted measurements against the hand-collected measurements taken from each cultivar treated with either 5 °C or 10 °C. Typical results for diverse pod morphotypes are illustrated in Fig. 5, including examples of cultivars with extreme pod lengths. These results show that the model's measurements correspond very well to manual measurements and the significance of each cultivar's treatment difference is also preserved despite the outputs for Zhongshuang II containing a number of false positives, poor quality masks and shorter measurements than its manual measured data, due to it is more extreme morphology (long and thin). Therefore, we conclude that the iterative active learning process has been effective. As shown in Fig. A.6, the valve length measurements based on DeepCanola accurately capture the genotypic differences and treatment effects across the entire BR17 dataset.

Fig. 6 shows p values calculated from a two-sample t-test denoting the statistical significance of the effect of treatment on its average pod length for each cultivar. This test was conducted for both hand-collected ground truth data and data predicted by DeepCanola. For $p < 0.01$ we have 100% concurrence between the two methods and for $p < 0.05$ we get a 95.7% concurrence between the two methods.

3.3. Results on disordered images

To evaluate how DeepCanola performs on real disordered image data and on previously unseen images, we utilised the BR9 dataset that has images from a similar earlier experiment where the pods had not been fully organised before imaging. Fig. 7 shows the qualitative results from two images of the dataset, typical of cases of more dense and more sparse scenes.

In situations where overlap occurs mainly perpendicularly and where density is relatively lower, the model is robust to disorder. However, where pods are largely orientated in parallel and almost touching or in a large dense group the model tends to become less accurate. A likely cause of this problem is the Mask R-CNN's two-stage architecture, where the model creates thousands of region proposals and suppresses the suboptimal ones via a Non-Max Suppression (NMS) algorithm. NMS treats the bounding boxes with high overlap as multiple proposals of one object, thus the algorithm works well for sparser scenes where juxtaposed multiple proposals most likely represent a single object. However, when objects are physically close or overlapping, the clusters of proposals merge and the NMS algorithm suppresses samples that it considers to make up a set of one object, but include multiple separate objects. The objects may be fused, one object may get suppressed, or a sub-optimal box including all or parts of multiple objects gets selected. In all such cases, the likelihood of correct detection and segmentation is reduced.

Results on dataset BR9 highlight the impact of ordering. Comparing predicted results to ground-truth masks with calculated valve lengths (see Section 2.4), we see a higher accuracy for ordered pods vs. disordered, with valve length R^2 = 0.9966, area R^2 = 0.9728 for ordered, and length R^2 = 0.9597, area R^2 = 0.9400 for disordered (see Fig. A.5). While physical ordering improves accuracy, there is a trade-off in terms of increased sample handling time, and as such would need to be assessed for each application. While highly dense and disordered images present challenges, the model still performs well on sparser samples. This suggests its potential applicability to similar images of related species.

3.4. Assessing the extent of generalisation

Although DeepCanola was designed specifically for highly constrained images of Brassica pod valves, we assessed how well the model had started to generalise, checking to what extent it could recognise the shape of pods in more diverse species and environments. Fig. 7 illustrates a panel of qualitative outputs — including samples of ordered and disordered BR11, Brassica stem material with attached pods, and related species in either ordered scenes or in a more natural but visually complex context. We see that, as expected, the model performs best on ordered samples of the same species as the training data, imaged under similar conditions. Remarkably, in similar imaging domains, the model is capable of detecting the valve region of pods in related species within the Brassicaceae family. We see examples of the morphologically diverse garlic mustard and radish fruits, with pedicel and beak attached, where the model has successfully detected and segmented the valve region from other pod parts. There are even some successful detections within in-field images of Brassica relatives, with partial detection of siliques. This supports our idea that the semi-synthetic training data allows the model to recognise similar objects under diverse conditions. Further improvement would be required to extract high-quality data, necessary for genetic or agronomic studies, from these more complex images. However, DeepCanola's ability to recognise morphologically diverse objects in these different domains highlights its potential for transfer learning on downstream tasks.

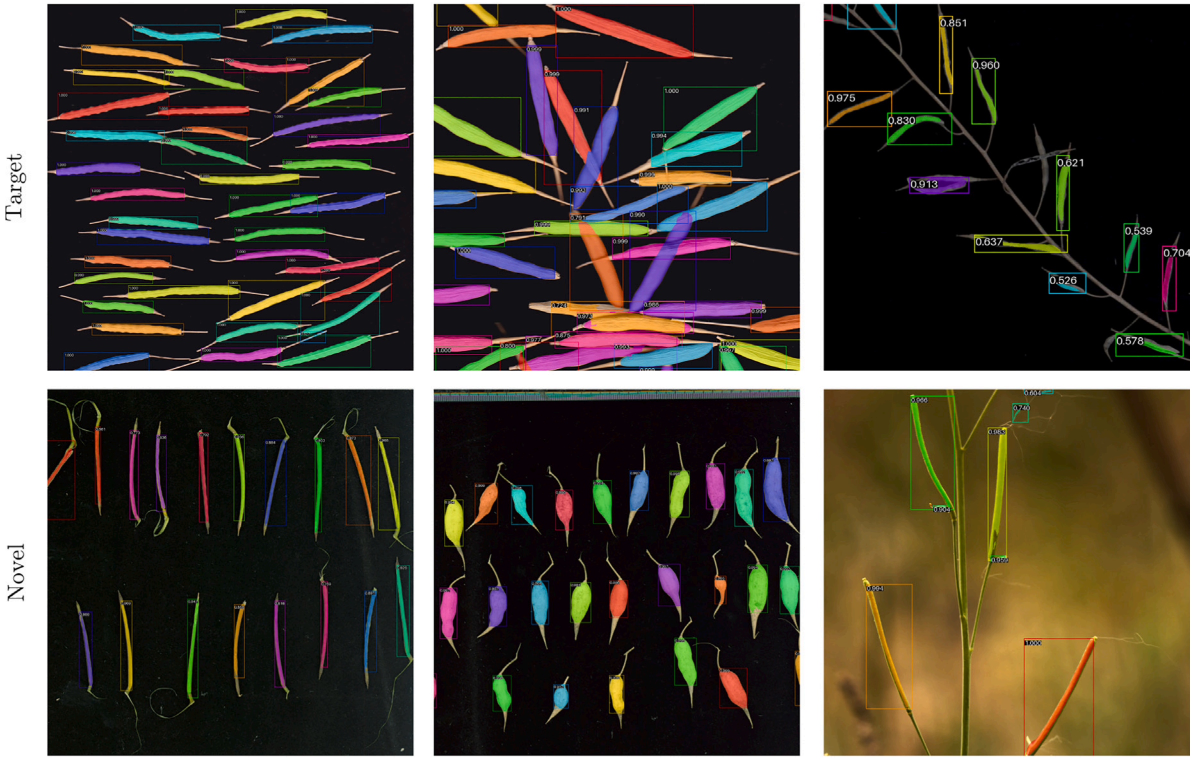


Fig. 7. DeepCanola generalisation capability — Examples test images showing samples from varying domains from the target and novel species. The top row of panels represents ordered, disordered, and different domain example images on the same species as used for model training. The bottom row represents ordered and different domain example images from different species (from left, *Alliaria petiolate* (Garlic mustard), *Raphanus raphanistrum* subsp. *sativus* (radish) and Arabidopsis). Colours are for visualisation purposes only, representing each instance of a pod valve detected by the model, and is used for both the instance's bounding box and segmentation. Note that colours are cycled through and reused for multiple instances.

4. Discussion

4.1. Overview

Here, we have iteratively trained a Mask R-CNN model to recognise and measure valves from scanned images of pods from diverse species and under different imaging scenarios. Object identification based on bounding boxes is useful for counting discrete botanical features such as fruit (Yang et al., 2021; Hamidinekoo et al., 2020), leaves, shoots (Wu et al., 2023b), wheat ears but such models often lack the spatial precision necessary for accurate phenotyping. Segmenting a spatially defined region (the valve or seed bearing region) within the bounding box of a larger more complex object (the pod) provides the precision required for accurate and flexible measurements and further extends the model's capabilities. Finally, we combine our model's predictions with classical computer vision methods to extract the pod curve, enabling accurate measurement of phenotypic traits such as valve length and valve area.

Our hybrid approach, combining human-in-the-loop active learning with semi-synthetic training data, significantly improved model performance and robustness across diverse imaging modalities, while dramatically reducing the burden of manual image annotation. Using an experimental dataset that was associated with detailed manual measurements of pod valves, the deep learning model performed as well as an expert human with an R^2 of 0.9930. The pod-valve annotations took about 39 person-hours in total to collect, whereas generation of the 1000-sample/44823-annotations (Dataset 4) took approximately one hour on a desktop workstation (hardware details in Section 2.3.3). The most rapid manual annotation was approximately 45 pods per hour, at which rate Dataset 4 would have taken approximately 996 person-hours to complete. This represents a 1000-fold increase in training data collection speed. The semi-synthetic dataset generation script is highly scalable, parallelised to run across multiple processor threads and could be easily adapted to other tasks.

Pod number and pod size are 2 key yield components for oilseed rape (Siles et al., 2021). Siles and co-workers examined 96 *B. napus* cultivars and found that the 2 main oilseed crop types (Winter and Spring OSR) share a common reproductive strategy with high numbers of long pods on the main inflorescence being the principal source of seed yield. Moreover, the relationship was further defined as being between valve length and the number of seeds per pod, which was exponential up to 5 cm and then linear. This motivated us to develop a method that could measure the valve region rather than the total pod length. However, initial versions performed poorly on cultivars with very long pods. Therefore, we developed an active learning process that increased the range of pod size that the model can accommodate by correctly labelling the pods that contributed to poor performance. The improved model then evaluated the effect of environmental conditions (2 types of simulated winter) with equal accuracy to an experienced crop scientist. Furthermore, we found that the model could be applied to additional datasets not included in the training, including other species, entire branches with attached pods and even images downloaded from the web.

4.2. Generalisation

We found that Thompson-inspired data augmentation allowed the model to generalise not only to related cultivars, but also to related species. Fig. 7 shows examples of *Raphanus raphanistrum* successfully segmented, although it is morphologically diverse from the brassica. We also see limited success in the presence of stem material. This suggests that the Thompson-inspired mathematical augmentation of pod shape has made the model robust in recognising similar shapes, even when potentially confusing material, such as stems, is present in the image. It also recognises valves in species with quite distinct overall pod morphologies. However, the precision of pod valve segmentation of these alternative species was marginally reduced compared to *B. napus*.

In some species, such as Arabidopsis, the beak-like region is very small relative to the rest of the fruit. However, inspection of Arabidopsis images labelled by DeepCanola indicates accurate restriction to the valve regions. Moreover, the shape of the beak region does not seem to impact the model's ability to accurately segment exclusively the valve region, which we see with unseen species such as the radish. This is likely due to the augmentation of pod shape variation, allowing the model to be robust to shape variation in these parts of the pod. The model also functions to a limited extent on even more extreme fruit shapes such as exhibited by Lunaria (honesty plant, Fig. A.4). However, further re-training would be required to achieve accurate results in the more morphologically distant samples.

We observed DeepCanola generalising well to diverse morphologies of pods of brassica and brassica relatives. To visualise the statistical distributions of the test images, t-distributed stochastic neighbour embedding (t-SNE (van der Maaten and Hinton, 2008)) is used to reduce the dimensionality of the feature maps from the hidden layers of the DeepCanola network, using both real-world test datasets and the test split of the generated dataset used for training. Fig. A.7 shows the t-SNE visualisation, where distinct clusters can be observed from each dataset and each pod ordering method. The semi-synthetic image cluster is positioned close to the other disordered datasets. This suggests that DeepCanola has learned that the generated and true disordered datasets have distinct yet similar compositions, and that the semi-synthetic data closely mimic the disordered images. We also extract the feature map outputs of the objects detected by our Mask R-CNN DeepCanola model. Fig. A.8 presents the t-SNE visualisation of each detected object across all test datasets. Due to the high overlap across datasets, we split each dataset and ordering method into its own sub-figure to better illustrate the distributions. Observing the disordered samples, we see that BR9 and BR11 pods occupy a similar region, while the generated samples lie within the same general area but exhibit more spread. This indicates that the objects created using the Thompson-inspired augmentation are similar to the true samples, but with greater morphological variation.

4.3. Limitations

A primary aim of our work was to reduce the time required to phenotype large quantities of brassica pods using image analysis. Reducing the need for personnel time input, our semi-synthetic data generation system quickly produces large quantities of high-quality, diverse (within the domain) training data. Despite the success of DeepCanola on uncluttered image data, the model has limitations. Some limitations are inherent in 2D images, such as overlap or occlusion. Other limitations stem from the nature of the semi-synthetic images, and the constrained domain of the images we generate.

A major component of the manual time requirement for phenotyping large quantities of pods is arranging samples. Therefore, we trained our model on samples that had varying degrees of complexity and order, resulting in both partial and fragmented occlusion. Partial occlusion is where a small region of the object is occluded. Fragmented occlusion is where multiple parts of the object are occluded, making it difficult for a model to determine which part belongs to what object. Examples of fragmented or partial occlusion can be seen in Fig. 3. Typically, currently available object detection models tend to fail when dealing with fragmented occlusion, including Mask R-CNN (Pegoraro and Pflugfelder, 2020; Pflugfelder and Auer, 2021). Although the model improved through the active learning process, when there are high levels of fragmented occlusion, there is simply not enough information within an image to accurately and consistently denote one object from another. We observe similar issues in dense clusters of pods, especially when physically touching and lying in parallel. Finding the optimal balance between order and disorder during image acquisition will therefore depend on the specific phenotyping application. This is a general limitation of single-image photogrammetry, but the use of disordered training data, both real and synthetic, means that DeepCanola has improved capabilities as a high-throughput phenotyping tool.

Complex or variable backgrounds, such as the image of Arabidopsis in Fig. 7 present several challenges, including but not limited to occlusion and out-of-focus objects, which may or may not be part of the subject of interest. These data types tend to generate more false positives and false negatives, which is perhaps not surprising since the model was not trained on such data. Although the current model is insufficiently robust for extracting accurate quantitative information from in-field images, generative AI and 3D modelling could potentially be used to produce new synthetic data to capture more diverse domains, including variable sample angles, lighting conditions, and backgrounds.

A potential issue with semi-synthetic data generation could also be the amplification of incorrectly labelled samples. If incorrectly labelled samples are frequent within the initial training data and then augmented to create additional abnormal training samples, this would cause the model to learn the incorrectly labelled relationship across different morphologies and locations. The training of object detection models benefits from having a highly informative and robust loss calculation that is a combination of all samples within a scene (Atkins et al., 2024). However, resampling and recomposition of the same object during augmentation may train the model to learn unintended relationships, such as unrealistic compositions that do not generalise well to real-world data with more constraints.

4.4. Addressing limitations

Object placement and density affect how well deep learning models detect objects. However, it can be challenging to find an optimum balance between the best possible images with minimal occlusion, while ensuring efficient image capture and accurate information extraction. For post-harvest material, minor changes in object arrangement can assist with detection within dense scenes. Three main features seem to improve valve detection: first, both ends of a given fruit should be visible, second, the obfuscated area along the body of the fruit should be minimised, and third, limiting the presence of dense clusters by having space between fruits. The DeepCanola, however, no longer requires careful arrangement of pods and tolerates disorder quite well. Generally, in situations of perpendicular overlap, DeepCanola confidently and correctly detects and segments both overlapping pods. However, when there are multiple pods overlapping one another in close proximity and with large areas obfuscated, the model tends to confuse each instance of a pod, returning some poor segmentations and detections. A simple approach to counter this effect, without manual arrangement of samples, is to limit the number of objects within a given image to reduce the chances of overlap. The relationship between object density and accuracy is also observed in commercial cereal grain-scanning systems, such as MARVIN, and has inspired the development of sophisticated grain-by-grain analysis systems such as C-grain Evershed et al. (2024). Being able to quantify the level of obfuscation within an image may also assist with understanding the limitations of deep learning models in dense scenes. A simple approach would be to quantify the number of fruits per square unit of area, calculating the relative density of fruits. Another measure could be the obfuscated area, quantifying the area of fruit within the image that is obfuscated; this would be difficult to measure for real-world images, however, for semi-synthetic images it could be found programmatically. Understanding how variation in these metrics affects outputted measurements, such as length, would help practitioners understand the degree of disorder that can be tolerated by deep learning models.

Deep learning models that use box-based non-max suppression (NMS), can perform poorly in dense scenes. However, some alternative algorithm choices may improve results. Bodla et al. (2017) extended the NMS algorithm to decay, instead of suppress, surrounding boxes. Named Soft-NMS, the algorithm decays the overlapping boxes' confidence scores instead of removing them outright. Decaying instead of suppressing boxes has the effect of allowing overlapping boxes with true high-confidence scores to remain, while removing other lower quality detections. This small change to the NMS algorithm has improved results in situations of overlapping objects in the COCO dataset. Similarly, Shepley et al. (2020) created Confluence, an alternative to NMS that uses the Manhattan distance between corners of boxes to decide which boxes in a set are likely to represent the same object. These alternative algorithms may assist with object detection in dense scenes. However, their utility must be assessed with the type of datasets we present in this work. Furthermore, using non-NMS-based deep learning models may improve results. For example, Liu et al. (2025) developed a Transformer-based model that could successfully detect green apples on the tree, including in situations where they were partly occluded by other fruit or by leaves and branches. Transformer-based segmentation models such as Mask2Former (Fu et al., 2019) may yield improved results. However, such models may introduce other forms of error when dealing with the task of dense object segmentation.

Deep learning models, ideally, should be robust to background and lighting variation, as this extends their utility. While consistent imaging conditions are easily implemented within a given laboratory, transfer between different sites often presents challenges, especially to real world situations. Re-training DeepCanola to be robust to background variation should be possible using our semi-synthetic data generation system. Novel natural backgrounds could be added to the background pool with negative samples, such as stripped (podless) stem material, placed within the image. This approach could be very useful for in-field studies where backgrounds are not easily controlled. Introducing a range of lighting conditions (i.e. directional illumination, variable shadows) could be more challenging in semi-synthetic setups. Gamma values can be augmented to vary the lighting intensity through post-processing, such as what was introduced for the background augmentation in the fourth round of training. However, gamma value augmentation changes can only mimic changes in intensity, not directionality of lighting. Generative AI systems can transfer lighting conditions from a target image to a source image (Xing et al., 2024). Such a system could be adapted to augment the lighting of a real or a generated image, allowing a model to become robust to diverse lighting conditions.

Handling high-density occlusions and variable imaging conditions in 2D image-based phenotyping remains a significant challenge. To ensure the quality of phenotypic measurements, statistical phenotypic metrics and post-processing techniques can help filter out problematic masks, thereby improving segmentation accuracy (Atkins et al., 2024). A potential extension of our approach is to enhance the pod detection models by incorporating an additional prediction head to estimate total fruit sizes in the image. Training the enhanced model would leverage our semi-synthetic data generation system to mitigate distortions caused by occlusions, varying shooting angles, and diverse phenotyping conditions.

4.5. Architecture selection

Mask R-CNN was selected as the architecture for our instance segmentation model due to its proven track record in creating quality outputs for non-real-time applications. The current state-of-the-art for instance segmentation on the COCO dataset is a large vision model based on Cascade Mask R-CNN (Fang et al., 2022). Cascade Mask R-CNN is an improvement over Mask R-CNN using a cascading architecture (Cai and Vasconcelos, 2021). The concept of cascade revolves around iterative cascading improvements from coarse to fine quality levels and is a well-tested system for creating quality segmentations. However, the main limitation of the DeepCanola model is object confusion in dense disordered scenes. This confusion can, at least partially, be attributed to non-max suppression (NMS) collapsing detections of multiple objects into a single output, and Cascade Mask R-CNN still relies upon NMS for its post-processing. Therefore, using the improved architecture would not address DeepCanola's primary limitation. Some transformer architectures, such as DETR (Carion et al., 2020) and its relatives, do not rely upon NMS for post-processing and potentially could handle dense scenes. However, transformer architectures require much more training data than their CNN counterparts (Gu et al., 2022) and are much more computationally complex than our proposed model, requiring more memory and computing to be able to train effectively.

5. Conclusion

This study proposes an active learning and semi-synthetic data generation system, training deep learning models to detect the yield-relevant valve region of brassica pods in both ordered and disordered scenes. The final model, named DeepCanola, shows good performance in segmenting pod valves, with the derived length phenotypic data showing a strong correlation with hand-collected data in both ordered ($R^2 = 0.9930$ and 0.9966) and disordered ($R^2 = 0.9597$) scenes. This indicates that high-quality phenotyping models can be created using semi-synthetic training data. We also demonstrate that applying simple mathematical transformations to the fruits assists the model's generalisation performance towards different cultivars, to related species and even to attached fruits in more complex scenarios.

CRediT authorship contribution statement

Larissa J.J. van Vliet: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Kieran Atkins:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Smita Kurup:** Writing – review & editing, Resources, Methodology, Investigation, Data curation. **Laura Siles:** Writing – review & editing, Resources, Methodology, Investigation, Data curation. **Jo Hepworth:** Writing – review & editing, Resources, Methodology, Investigation, Data curation. **Fiona M.K. Corke:** Writing – review & editing, Resources, Methodology, Investigation, Data curation. **John H. Doonan:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization. **Chuan Lu:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Funding

The authors acknowledge financial support from Biotechnology and Biological Sciences Research Council grants: Brassica Rapeseed And Vegetable Optimisation strategic Longer and Larger fund (BRAVO sLOLA) (BB/P003095/1), Institute Strategic Programme 'Genes in the environment' (BB/P013511/1) and 'Tailoring Plant Metabolism' (BB /P012633/1) and BBSRC-NCG in Plant Phenotyping (BBS/E/W/00128 44A); and EPSRC grant - Miscanthus AI- Plant selection and breeding for Net Zero (EP/Y005430/1). KA was supported by UKRI BBSRC FoodBioSystems Doctoral Training Partnership (DTP), grant number BB/T008776/1.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank NPPC Staff for growing the plant populations, Richard Pickering for his assistance with pod harvesting, and Gabi Clifton-Brown and Karen Askew for helping with image annotation.

Appendix. Supplementary

See Figs. A.1–A.8 and Table A.1.

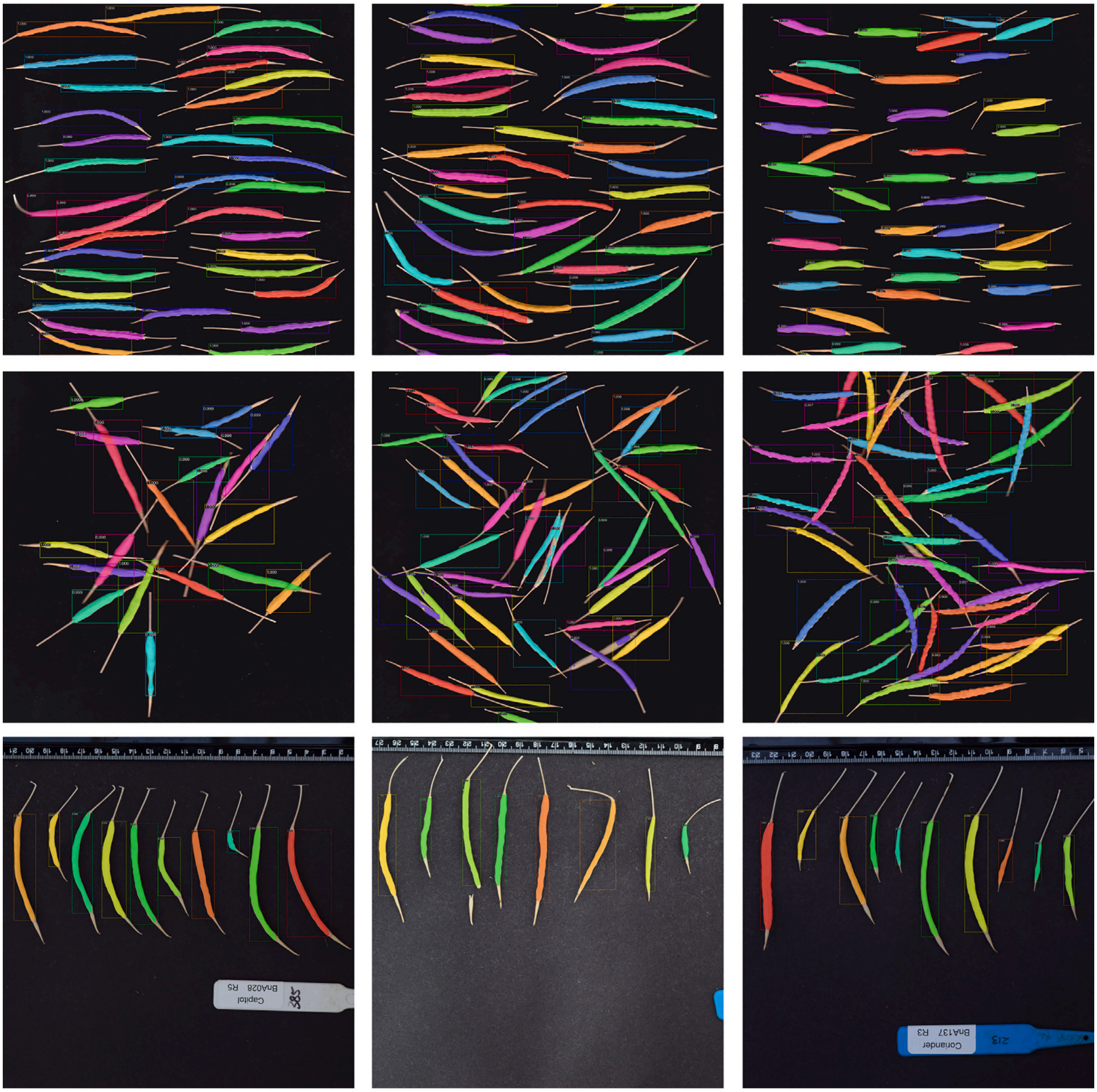
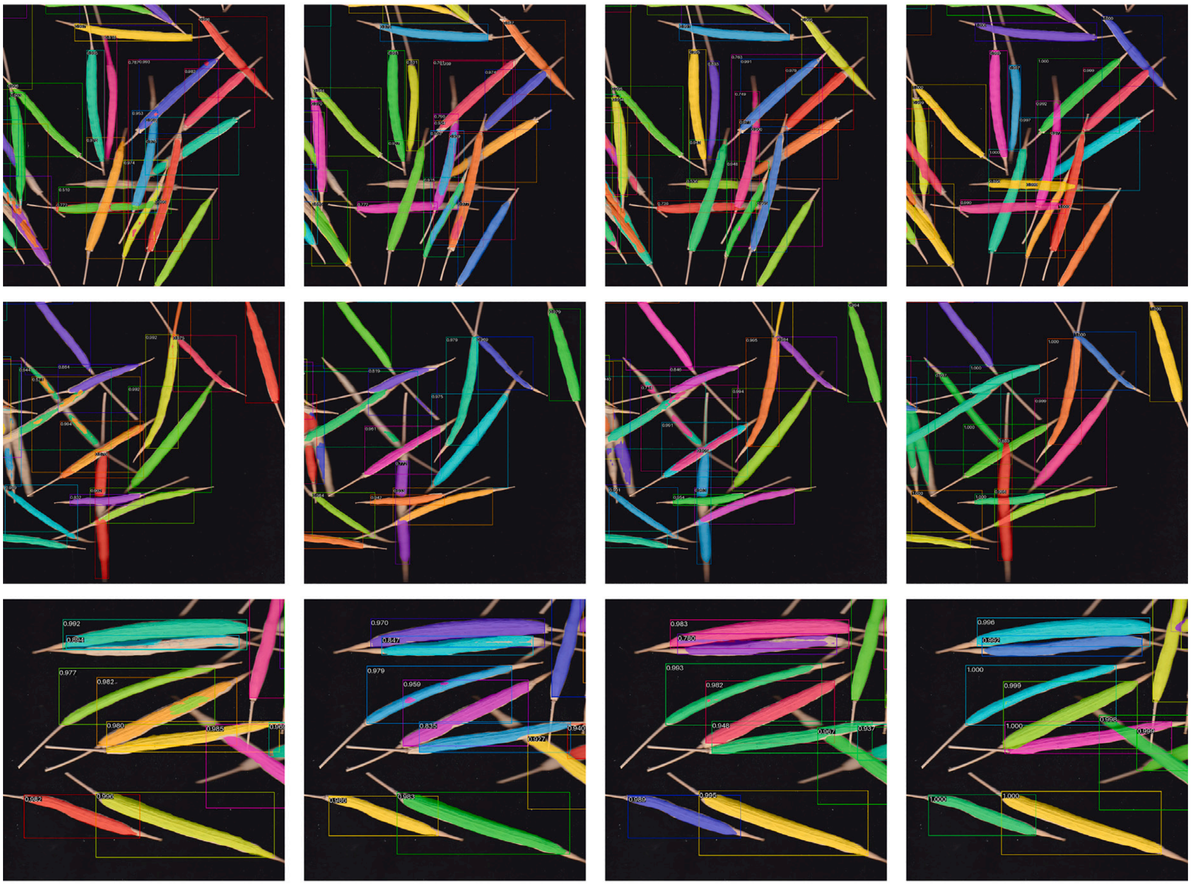


Fig. A.1. Typical DeepCanola Outputs — Typical examples from DeepCanola on both ordered and disordered images from both datasets BR9 and BR17 (rows 1 & 2), alongside a dataset of *Brassica napus* supplied by Rothamsted Research (row 3), which was collected on a distant site and not for the purpose of being used as a validation example for this system. Colours are for visualisation purposes only, representing each instance of a pod valve detected by the model, and is used for both the instance's bounding box and segmentation. Note that colours are cycled through and reused for multiple instances.

Table A.1
Model hyperparameters used for training.

Hyperparameters	Values
Backbone	ResNet-50-FPN
Input size	(2631 x 1914 x 3)
Learning rate	0.005
Momentum	0.9
Weight decay	0.0005
LR scheduler gamma	0.1
LR scheduler epoch steps	8, 11
Total training epochs	12



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

Fig. A.2. Extra active learning progression examples — Example patch of test image from disordered BR9 dataset showing results from each of the step of the active learning process, expanding on Fig. 3. Colours are for visualisation purposes only, representing each instance of a pod valve detected by the model, and is used for both the instance's bounding box and segmentation. Note that colours are cycled through and reused for multiple instances.

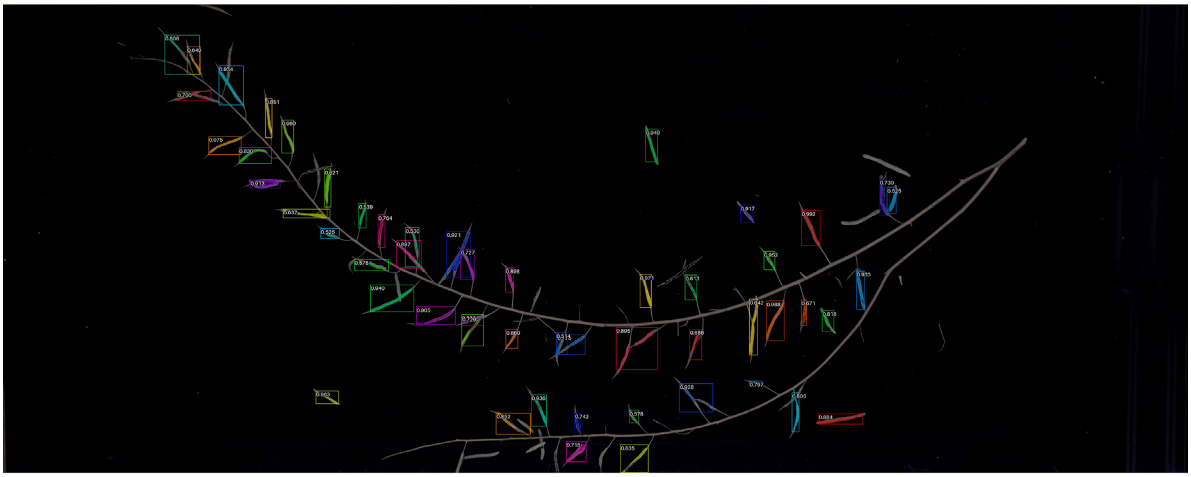


Fig. A.3. Example DeepCanola outputs on Brassica stem material — Visualised outputs of DeepCanola on example image of *Brassica napus* stem material with pods attached illustrating both generalisation and limitations of DeepCanola. Colours are for visualisation purposes only, representing each instance of a pod valve detected by the model, and is used for both the instance's bounding box and segmentation. Note that colours are cycled through and reused for multiple instances.

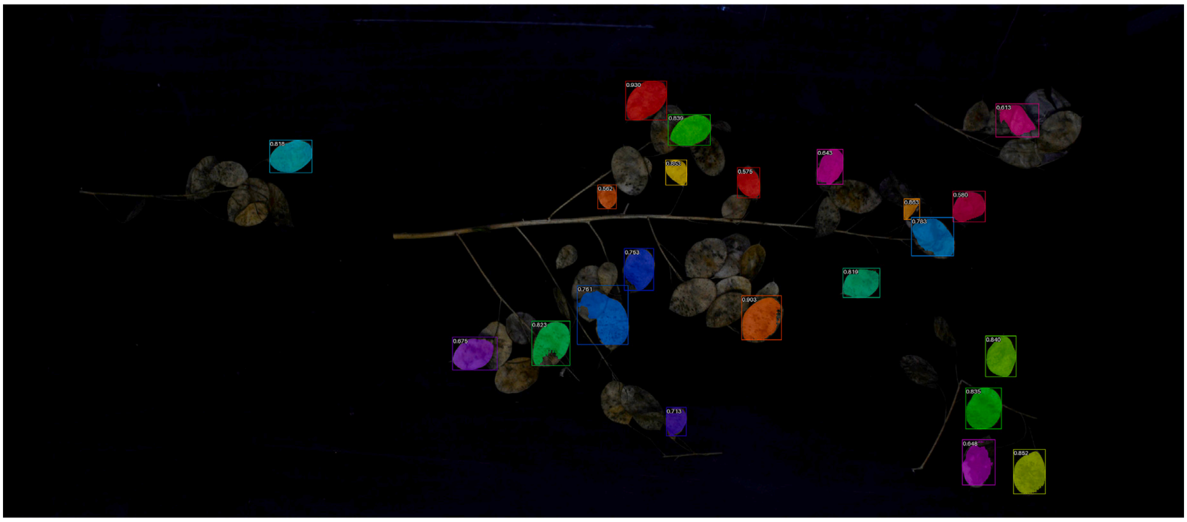


Fig. A.4. Example DeepCanola outputs on *Lunaraia annua* stem material — Visualised outputs of DeepCanola on example image of *Lunaraia annua* stem material with its highly morphologically diverse pods attached showing illustrating both generalisation and limitations of DeepCanola. Colours are for visualisation purposes only. Colours are for visualisation purposes only, representing each instance of a pod valve detected by the model, and is used for both the instance's bounding box and segmentation. Note that colours are cycled through and reused for multiple instances.

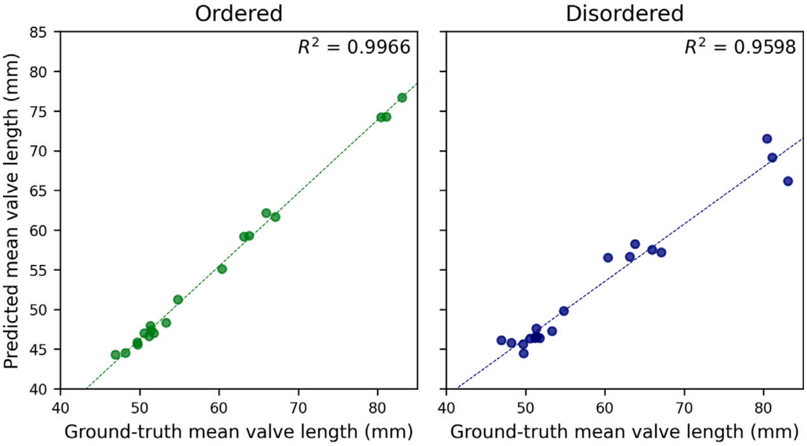


Fig. A.5. DeepCanola valve length results on BR9 dataset for ordered and disordered images — Scatter plot showing the calculated valve length from ground-truth masks against calculated valve length from DeepCanola. Shown for both ordered and disordered pods. R^2 value is shown in the upper right corner, along with a linear regression line. Ordered images are shown in green, and disordered images are shown in blue.

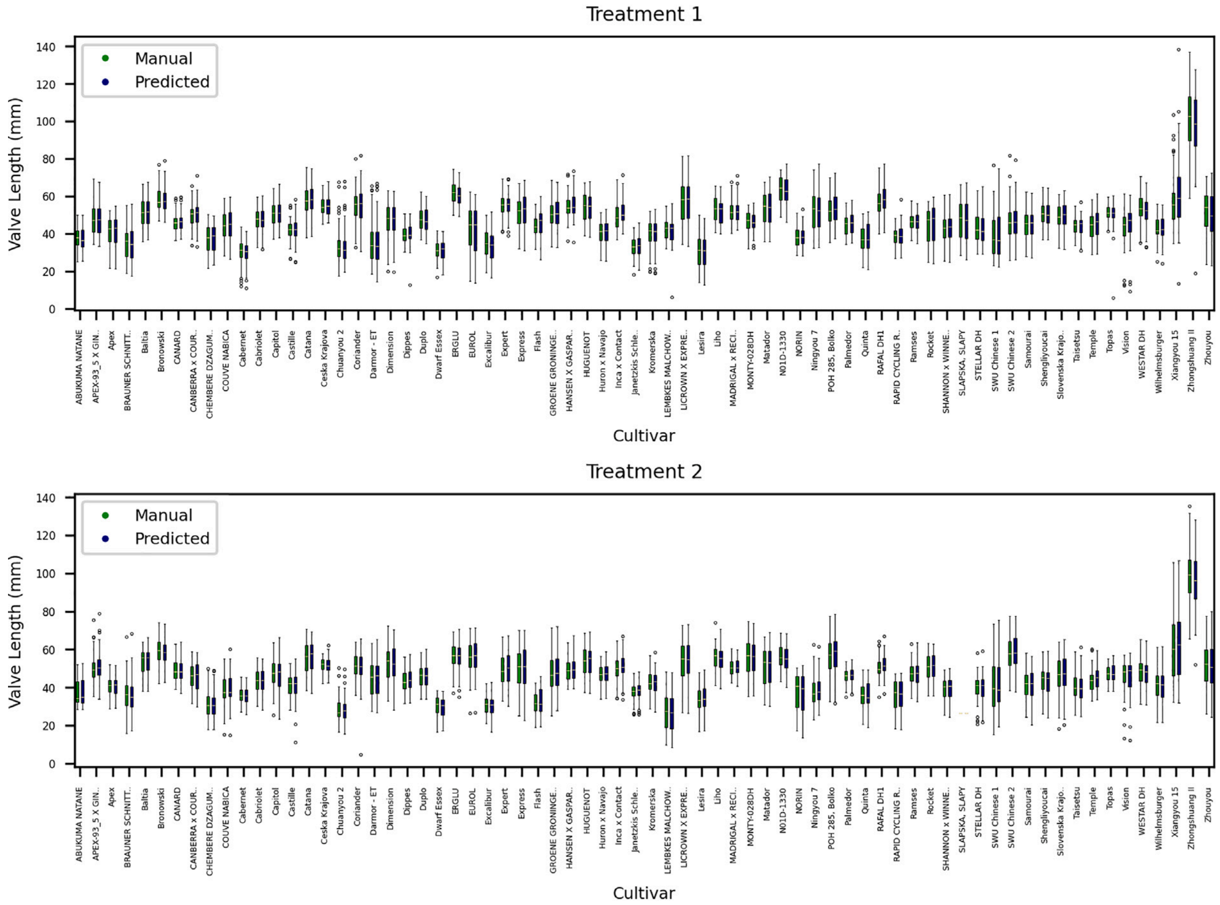


Fig. A.6. Manual and DeepCanola valve length measurements on BR17 — The results for the manually measured data (green — left box in group) and the predicted data (blue — right box in group) for treatment 1 (upper) and treatment 2 (lower) and cultivar in the BR17 dataset.

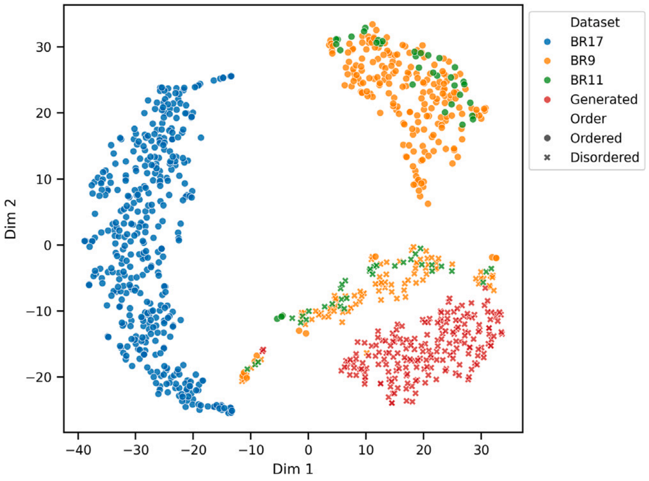


Fig. A.7. Image-level t-SNE visualisation of the feature maps produced by the model backbone on all test images from each dataset. Colours represent different datasets, dot style represents ordering of the pods within the image. Feature map extracted from the third (final) layer of the FPN subnetwork attached to the CNN backbone. t-SNE reduced 256-dimension feature map to 2-dimension output (Dim 1 & Dim 2). Here, *Generated dataset* is Dataset 4, the test-split of the dataset used to train DeepCanola. Parameters: perplexity = 30, iterations = 1000. Note that the overall shape and clustering shown in this figure was consistent for perplexity 10–50, and with higher perplexities we observed the real and generated disordered clusters begin to combine.

Data availability

Model weights, training and inference scripts, along with dataset generation code are available at https://github.com/kieranatkins/deepcanola.Pod image pool, valve annotation pool, background pool, generated training and testing datasets at each step of the active learning process, along with valve measurements on the BR17 dataset are provided under DOI: 10.5281/zenodo.14235543. Note that (iNaturalist community, 2023) is a public dataset and images used are not included in DOI.

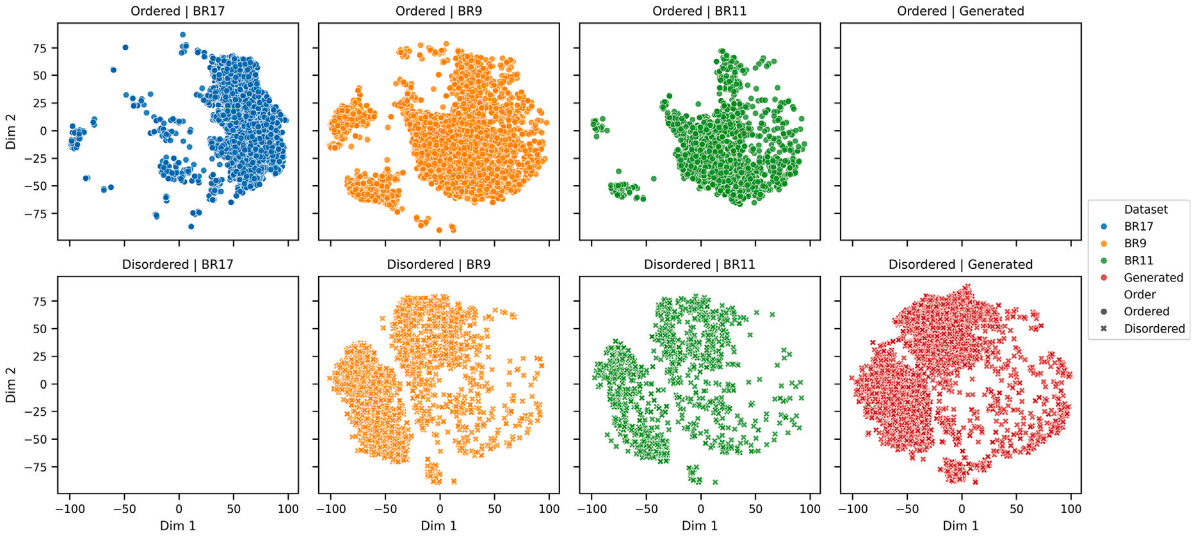


Fig. A.8. Object-level t-SNE visualisation of the feature maps produced by the model backbone on all test images from each dataset. Feature map extracted from cropped features inputted to box head of the Mask R-CNN network. t-SNE reduced 256-dimension feature map to 2-dimension output (Dim 1 & Dim 2). All sub-figures are from the same t-SNE embedding, however dataset and pod ordering types have been split for visibility. Colours/columns represent different datasets, rows represent different pod ordering. Here, *Generated dataset* is Dataset 4, the test-split of the dataset used to train DeepCanola. Parameters: perplexity = 30, iterations = 1000. Note that the overall shape and clustering shown within these figures was consistent for higher perplexity and iteration values.

References

2024. Domain targeted synthetic plant style transfer using stable diffusion, lora and controlnet. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. CVPRW, pp. 5375–5383. http://dx.doi.org/10.1109/CVPRW63382.2024.00546.

Atkins, K., Garzón-Martínez, G., Lloyd, A., Doonan, J.H., Lu, C., 2024. Unlocking the power of ai for phenotyping fruit morphology in arabidopsis. GigaScience http://dx.doi.org/10.1093/gigascience/giae123.

Blok, P.M., Kootstra, G., Elghor, H.E., Diallo, B., van Evert, F.K., van Henten, E.J., 2022. Active learning with maskal reduces annotation effort for training mask r-cnn on a broccoli dataset with visually similar classes. Comput. Electron. Agric. 197, 106917. http://dx.doi.org/10.1016/j.compag.2022.106917, URL: https://www.sciencedirect.com/science/article/pii/S0168169922002344.

Bodla, N., Singh, B., Chellappa, R., Davis, L.S., 2017. Improving object detection with one line of code. CoRR abs/1704.04503, URL: http://arxiv.org/abs/1704.04503, arXiv:1704.04503.

Cai, Z., Vasconcelos, N., 2021. Cascade r-cnn: High quality object detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 43, 1483–1498. http://dx.doi.org/10.1109/TPAMI.2019.2956516.

Calderwood, A., Lloyd, J., Tudor, E.H., Jones, D.M., Woodhouse, S., Bilham, L., Chinoy, C., Williams, K., Corke, F., Doonan, J.H., Ostergaard, L., Irwin, J.A., Wells, R., Morris, R.J., 2021. Total flc transcript dynamics from divergent paralogue expression explains flowering diversity in brassica napus. New Phytol. 229, 3534–3548. http://dx.doi.org/10.1111/nph.17131, URL: https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.17131.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. CoRR http://dx.doi.org/10.48550/arXiv.2005.12872, URL: https://arxiv.org/abs/2005.12872.

Chandra, A.L., Desai, S.V., Balasubramanian, V.N., Ninomiya, S., Guo, W., 2020. Active learning with point supervision for cost-effective panicle detection in cereal crops. Plant Methods 16, http://dx.doi.org/10.1186/s13007-020-00575-8, URL: http://dx.doi.org/10.1186/S13007-020-00575-8.

iNaturalist community, 2023. Observations of brassicaceae. (Accessed 12 July 2023). URL: https://www.inaturalist.org.

Diepenbrock, W., 2000. Yield analysis of winter oilseed rape (brassica napus l.): a review. Field Crop. Res. 67, 35–49. http://dx.doi.org/10.1016/S0378-4290(00)00082-4, URL: https://www.sciencedirect.com/science/article/pii/S0378429000000824.

Evershed, D., Durkan, E.J., Hasler, R., Corke, F., Doonan, J.H., Howarth, C.J., 2024. Critical evaluation of the cgrain value™ as a tool for rapid morphometric phenotyping of husked oat (avena sativa l.) grains. Seeds 3, 436–455. http://dx.doi.org/10.3390/seeds3030030, URL: https://www.mdpi.com/2674-1024/3/3/30.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y., 2022. Eva: Exploring the limits of masked visual representation learning at scale. http://dx.doi.org/10.48550/arXiv.2211.07636, URL: https://arxiv.org/abs/2211.07636.

Fu, C., Shvets, M., Berg, A.C., 2019. Retinamask: Learning to predict masks improves state-of-the-art single-shot detection for free. CoRR http://dx.doi.org/10.48550/arXiv.1901.03353, URL: http://arxiv.org/abs/1901.03353.

Granland, K., Newbury, R., Chen, Z., Ting, D., Chen, C., 2022. Detecting occluded y-shaped fruit tree segments using automated iterative training with minimal labeling effort. Comput. Electron. Agric. 194, 106747. http://dx.doi.org/10.1016/j.compag.2022.106747, URL: https://www.sciencedirect.com/science/article/pii/S0168169922000643.

Gu, W., Bai, S., Kong, L., 2022. A review on 2d instance segmentation based on deep neural networks. Image Vis. Comput. 120, 104401. http://dx.doi.org/10.1016/j.imavis.2022.104401, URL: https://www.sciencedirect.com/science/article/pii/S0262885622000300.

Gulden, R.H., Warwick, S.I., Thomas, A.G., 2008. The biology of canadian weeds. 137. brassica napus l. and b. rapa l.. Can. J. Plant Sci. 88, 951–996. http://dx.doi.org/10.4141/CJPS07203.

Hamidinekoo, A., Garzón-Martínez, G.A., Ghahremani, M., Corke, F.M.K., Zwiggelaar, R., Doonan, J.H., Lu, C., 2020. DeepPod: a convolutional neural network based quantification of fruit number in Arabidopsis. GigaScience 9, giaa012. http://dx.doi.org/10.1093/gigascience/giaa012.

He, K., Gkioxari, G., Dollár, P., Girshick, R.B., 2017. Mask R-CNN. CoRR http://dx.doi.org/10.48550/arXiv.1703.06870, URL: http://arxiv.org/abs/1703.06870.

Hossain, S., Kadkol, G., Raman, R., Salisbury, P., Raman, H., 2012. Breeding brassica napus for shatter resistance. In: Plant Breeding. IntechOpen, pp. 313–332. http://dx.doi.org/10.5772/29051.

Kumar, S., Luo, W., Kantor, G., Sycara, K.P., 2019. Active learning with gaussian processes for high throughput phenotyping. CoRR http://dx.doi.org/10.48550/arXiv.1901.06803, URL: http://arxiv.org/abs/1901.06803.

Łangowski, Ł., Stacey, N., Ostergaard, L., 2016. Diversification of fruit shape in the brassicaceae family. Plant Reprod. 29, http://dx.doi.org/10.1007/s00497-016-0278-6, URL: https://link.springer.com/article/10.1007/s00497-016-0278-6.

Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, C.L., 2014. Microsoft COCO: common objects in context. CoRR http://dx.doi.org/10.48550/arXiv.1405.0312, URL: http://arxiv.org/abs/1405.0312.

Liu, X., Hu, C., Li, P., Automatic segmentation of overlapped poplar seedling leaves combining mask r-cnn and dbscan. Computers and Electronics in Agriculture 178, 105753. http://dx.doi.org/10.1016/j.compag.2020.105753, URL: https://www.sciencedirect.com/science/article/pii/S0168169920311777.

Liu, Q., Meng, H., Zhao, R., Ma, X., Zhang, T., Jia, W., 2025. Green apple detector based on optimized deformable detection transformer. Agriculture 15, http://dx.doi.org/10.3390/agriculture15010075, URL: https://www.mdpi.com/2077-0472/15/1/75.

Lu, W., Du, R., Niu, P., Xing, G., Luo, H., Deng, Y., Shu, L., 2022. Soybean yield preharvest prediction based on bean pods and leaves image recognition using deep learning neural network combined with grnn. Front. Plant Sci. 12, http://dx.doi.org/10.3389/fpls.2021.791256, URL: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2021.791256.

van der Maaten, L., Hinton, G.E., 2008. Visualizing data using t-sne. J. Mach. Learn. Res. 9, 2579–2605, URL: https://api.semanticscholar.org/CorpusID:5855042.

Napier, C.C., Cook, D.M., Armstrong, L., Diepeveen, D., 2023. A synthetic wheat l-system to accurately detect and visualise wheat head anomalies. In: Proceedings of the 3rd International Conference on Smart and Innovative Agriculture (ICoSIA 2022). Atlantis Press, pp. 379–391. http://dx.doi.org/10.2991/978-94-6463-122-7_36.

Pegoraro, J., Pflugfelder, R.P., 2020. The problem of fragmented occlusion in object detection. CoRR http://dx.doi.org/10.48550/arXiv.2004.13076, URL: https://arxiv.org/abs/2004.13076.

Pflugfelder, R., Auer, J., 2021. Person localisation under fragmented occlusion. In: 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance. AVSS, pp. 1–8. http://dx.doi.org/10.1109/AVSS52988.2021.9663791, URL: https://ieeexplore.ieee.org/document/9663791. 2021 17th IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS) ; Conference date: 16-11-2021 Through 19-11-2021.

Rawat, S., Chandra, A.L., Desai, S.V., Balasubramanian, V.N., Ninomiya, S., Guo, W., 2022. How useful is image-based active learning for plant organ segmentation? Plant Phenomics http://dx.doi.org/10.34133/2022/9795275, URL: https://spj.science.org/doi/abs/10.34133/2022/9795275.

Ren, S., He, K., Girshick, R.B., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. CoRR http://dx.doi.org/10.48550/arXiv.1506.01497, URL: http://arxiv.org/abs/1506.01497.

Schiessl, S.V., Huettel, B., Kuehn, D., Reinhardt, R., Snowdon, R.J., 2017. Flowering time gene variation in brassica species shows evolutionary principles. Front. Plant Sci. 8, http://dx.doi.org/10.3389/fpls.2017.01742, URL: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2017.01742.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.Y., White, D., Hartenstein, V., Eliceiri, K., Tomancak, P., Cardona, A., 2012. Fiji: An open-source platform for biological-image analysis. Nature Methods 9, 676–682. http://dx.doi.org/10.1038/nmeth.2019, URL: https://www.nature.com/articles/nmeth.2019.

Shepley, A., Falzon, G., Kwan, P., 2020. Confluence: A robust non-iou alternative to non-maxima suppression in object detection. CoRR URL: https://arxiv.org/abs/2012.00257, arXiv:2012.00257.

Siles, L., Hassall, K.L., Sanchis Gritsch, C., Eastmond, P.J., Kurup, S., 2021. Uncovering trait associations resulting in maximal seed yield in winter and spring oilseed rape. Front. Plant Sci. 12, http://dx.doi.org/10.3389/fpls.2021.697576, URL: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2021.697576.

Su, W.H., Zhang, J., Yang, C., Page, R., Szinyei, T., Hirsch, C., Steffenson, B., 2020. Automatic evaluation of wheat resistance to fusarium head blight using dual mask-rcnn deep learning frameworks in computer vision. Remote. Sens. 13, 1–21. http://dx.doi.org/10.3390/rs13010026, URL: https://www.mdpi.com/2072-4292/13/1/26.

Thompson, D.W., 1961. On Growth and Form. Cambridge University Press, Cambridge, URL: https://www.gutenberg.org/ebooks/55264.

Toda, Y., Okura, F., Ito, J., Okada, S., Kinoshita, T., Tsuji, H., Saisho, D., 2020. Training instance segmentation neural network with synthetic datasets for crop seed phenotyping. Commun. Biology 3, 173. http://dx.doi.org/10.1038/s42003-020-0905-5.

Williams, K., Hepworth, J., Nichols, B.S., Corke, F., Woolfenden, H., Paajanen, P., Steuernagel, B., Østergaard, L., Morris, R.J., Doonan, J.H., Wells, R., 2023. Integrated phenomics and genomics reveals genetic loci associated with inflorescence growth in brassica napus. BioRxiv http://dx.doi.org/10.1101/2023.03.31.535149, URL: https://www.biorxiv.org/content/early/2023/04/04/2023.03.31.535149.

Wu, X., Fan, X., Luo, P., Choudhury, S.D., Tjahjadi, T., Hu, C., 2023b. From laboratory to field: Unsupervised domain adaptation for plant disease recognition in the wild. Plant Phenomics 5, 0038. http://dx.doi.org/10.34133/plantphenomics.0038, URL: https://spj.science.org/doi/abs/10.34133/plantphenomics.0038.

Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C., 2023a. Datasetdm: Synthesizing data with perception annotations using diffusion models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 54683–54695. http://dx.doi.org/10.48550/arXiv.2308.06160.

Xing, X., Groh, K., Karaoglu, S., Gevers, T., Bhattad, A., 2024. Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. http://dx.doi.org/10.48550/arXiv.2412.00177, arXiv:2412.00177.

Yang, S., Zheng, L., Yang, M., Wu, T., Sun, S., Tomasetto, F., Wang, M., 2021. A synthetic datasets based instance segmentation network for high-throughput soybean pods phenotype investigation. Expert Syst. Appl. 192, 116403. http://dx.doi.org/10.1016/j.eswa.2021.116403.