frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 14 April 2023
DOI 10.3389/fpls.2023.1108355

OPEN ACCESS

EDITED BY
Zhanyou Xu,
United States Department of Agriculture (USDA), United States

REVIEWED BY
Jun Yan,
China Agricultural University, China
Yong Suk Chung,
Jeju National University, Republic of Korea

*CORRESPONDENCE
Talukder Z. Jubery
znjubery@iastate.edu
Baskar Ganapathysubramanian
baskarg@iastate.edu

SPECIALTY SECTION
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

RECEIVED 25 November 2022
ACCEPTED 28 March 2023
PUBLISHED 14 April 2023

CITATION
Dong D, Nagasubramanian K, Wang R, Frei UK, Jubery TZ, Lübberstedt T and Ganapathysubramanian B (2023) Self-supervised maize kernel classification and segmentation for embryo identification. *Front. Plant Sci.* 14:1108355. doi: 10.3389/fpls.2023.1108355

Self-supervised maize kernel classification and segmentation for embryo identification

David Dong[1,2], Koushik Nagasubramanian[2,3], Ruidong Wang[4], Ursula K. Frei[4], Talukder Z. Jubery[2,5]*, Thomas Lübberstedt[4] and Baskar Ganapathysubramanian[2,3,5]*

[1]Ames High School, Ames, IA, United States, [2]Translational AI Center, Iowa State University, Ames, IA, United States, [3]Department of Electrical Engineering, Iowa State University, Ames, IA, United States, [4]Department of Agronomy, Iowa State University, Ames, IA, United States, [5]Department of Mechanical Engineering, Iowa State University, Ames, IA, United States

Introduction: Computer vision and deep learning (DL) techniques have succeeded in a wide range of diverse fields. Recently, these techniques have been successfully deployed in plant science applications to address food security, productivity, and environmental sustainability problems for a growing global population. However, training these DL models often necessitates the large-scale manual annotation of data which frequently becomes a tedious and time-and-resource- intensive process. Recent advances in self-supervised learning (SSL) methods have proven instrumental in overcoming these obstacles, using purely unlabeled datasets to pre-train DL models.

Methods: Here, we implement the popular self-supervised contrastive learning methods of NNCLR Nearest neighbor Contrastive Learning of visual Representations) and SimCLR (Simple framework for Contrastive Learning of visual Representations) for the classification of spatial orientation and segmentation of embryos of maize kernels. Maize kernels are imaged using a commercial high-throughput imaging system. This image data is often used in multiple downstream applications across both production and breeding applications, for instance, sorting for oil content based on segmenting and quantifying the scutellum's size and for classifying haploid and diploid kernels.

Results and discussion: We show that in both classification and segmentation problems, SSL techniques outperform their purely supervised transfer learning-based counterparts and are significantly more annotation efficient. Additionally, we show that a single SSL pre-trained model can be efficiently finetuned for both classification and segmentation, indicating good transferability across multiple downstream applications. Segmentation models with SSL-pretrained backbones produce DICE similarity coefficients of 0.81, higher than the 0.78 and 0.73 of

those with ImageNet-pretrained and randomly initialized backbones, respectively. We observe that finetuning classification and segmentation models on as little as 1% annotation produces competitive results. These results show SSL provides a meaningful step forward in data efficiency with agricultural deep learning and computer vision.

KEYWORDS

self-supervised, classification, embryo identification, segmentation, high-throughput sorting

1 Introduction

Deep learning (DL) for computer vision applications has recently become a boon to innovations in agricultural efficiency. These methods have transformed how we extract various agronomically relevant plant traits under laboratory and field conditions (Fahlgren et al., 2015; Ubbens and Stavness, 2017; Singh et al., 2018; Guo et al., 2021). Automatically and rapidly extracting plant traits can be a game-changer in terms of reducing food costs and improving production efficiencies, improving sustainability by reducing waste, and providing a better understanding of adapting crops for climate change. Deep learning methods have been used in various agricultural applications to identify, classify, quantify, and predict traits (Mohanty et al., 2016; Naik et al., 2017; Pound et al., 2017; Dobrescu et al., 2019; Jubery et al., 2021). With the availability of high-throughput data acquisition tools that produce large amounts of good-quality data, the major bottleneck in deploying DL-based computer vision tools is the need for large amounts of labeled data to train these DL models. Data annotation or labeling is the main development barrier to building high-quality DL models, especially since labeling the raw data often requires domain experts to annotate images. Data annotation by an expert with domain-specific knowledge is a tedious and expensive task. The DL community is exploring various strategies to break this dependency on a large quantity of annotated data to train DL models in a label-efficient manner, including approaches like active learning (Nagasubramanian et al., 2021), transfer learning (Jiang and Li, 2020), weakly supervised learning (Ghosal et al., 2019; Körschens et al., 2021) and the more recent advances in self-supervised learning (Jing and Tian, 2020; Marin Zapata et al., 2021; Nagasubramanian et al., 2022). Transfer learning has been widely utilized in plant phenomics applications for classification and segmentation tasks (Wang et al., 2019; Kattenborn et al., 2021). Recently, self-supervised learning has been applied to improve classification and segmentation models (Güldenring and Nalpantidis, 2021; Nagasubramanian et al., 2022; Lin et al., 2023). In this work, we focus on deploying self-supervised learning approaches to the problem of characterizing maize kernels that are imaged in a commercial high-throughput seed imaging system [Qsorter technologies (QualySense)]. We consider two vision tasks – first, identify if the maize kernels are correctly oriented for downstream analysis (a classification task), and second, segment out the kernel scutellum from the correctly oriented seeds (a segmentation task).

The ability to accurately and efficiently segment maize kernel scutellum has significant utility for both production and breeding application. Maize oil (corn oil) is extracted from corn kernels through milling (Paulsen and Hill, 1985). Milling processes are integrated into the production of corn starch, sugar, syrup, alcohol, and byproducts like gluten feed, along with corn oil. Of the 1.1 billion metric tons of corn produced annually around the world, over 3.5 million are used for oil production (Ward and Singh, 2002; Lee et al., 2021). Almost all oil is found in the embryo of the kernel (Paulsen and Hill, 1985). The ability to sort seeds for embryo/scutellum size is a significant value addition. Similarly, the non-destructive sorting of single seeds based on oil content (OC) has been shown to be useful for early-generation screening to improve the efficiency of breeding (Silvela et al., 1989; Xu et al., 2019) and for haploid selection in an oil-inducer-based doubled haploid breeding program (Chaikam et al., 2019; Aboobucker et al., 2022). Over the past few years, nuclear magnetic resonance (NMR) (Melchinger et al., 2017; Yang et al., 2018), fluorescence imaging (Boote et al., 2016), near-infrared (NIR) reflectance spectroscopy (Jiang et al., 2007; Armstrong et al., 2011; Jones et al., 2012; Gustin et al., 2020), hyperspectral imaging (Weinstock et al., 2006), and line-scan Raman hyperspectral imaging (Liu et al., 2022) have been developed to measure or predict oil content. However, these methods and tools are expensive. On the other hand, sorting based on NIR reflectance is less costly, has been around for a long time (McClure, 2003; Halcro et al., 2020), and has worked well to predict protein and starch content. However, using those tools to measure OC is not easy because the position of the embryo/scutellum to the camera/light source (Spielbauer et al., 2009) strongly affects OC measurements of single seeds, which leads to significant standard errors. Several currently available NIR spectra-based high throughput single seed sorting devices capture RGB images of the seed along with the NIR spectrum (QualySense; Satake-USA). These images can be used to identify the correctly orientated seed and quantify the relative size of the embryo to the seed, which, coupled with the NIR spectrum, could be used to improve the prediction of OC.

This work aims to design an end-to-end DL framework that classifies kernels based on their orientation and segments the embryos of correctly oriented kernels. Accurately performing these steps will allow us to, in the future, predict corn OC with high accuracy. Figure 1 illustrates this pipeline. A challenge in accomplishing this goal is that DL techniques often rely on having access to large datasets of annotated images for successful training results. This problem motivates our approach of using self-supervised contrastive. The self-supervised pretraining procedure automatically uses unlabeled data to generate pretrained labels (Misra and Maaten, 2020). It does so by solving a pretext task suited for learning representations, which in computer vision typically consists of learning invariance to image augmentations like rotation and color transforms, producing feature representations that ideally can be easily adapted for use in a downstream task. After obtaining this pre-trained model, we apply standard DL to finetune the model with a smaller labeled dataset. The smaller labeled dataset is used to reduce the effect of possible inaccuracies in the pseudo-labels from the self-supervised task (Zhai et al., 2019). The orientation of corn kernels must maintain consistency between measurements and be oriented to fully display the embryo. The goal of the segmentation problem is then to identify the embryo amidst the background and the rest of each kernel.

Our contributions in this paper are 1) the creation of an end-to-end DL pipeline for kernel classification and segmentation, facilitating downstream applications in OC prediction, 2) to assess capabilities of self-supervised learning regarding annotation efficiency, and 3) illustrating the ability of self-supervised pretraining to create models that can be finetuned for diverse downstream applications. Beyond the direct application of the classification and segmentation capabilities of the learned representations, using self-supervised techniques, in general, could accelerate the development of computer vision techniques for ag applications, skipping several stages of arduous and time-consuming data collection.

2 Materials and methods

2.1 Dataset

2.1.1 Dataset for classification by imaging orientation

The classification dataset consists of 44,286 RGB 492-pixel by 240-pixel images of maize kernels of various accessions taken using the RGB imaging tools of QSorter. Of these, 2697 were manually labeled into two classes: “oriented” and “non-oriented.” Kernels that belong to the “oriented” class were deemed appropriate for calculating internal OC within the embryo/germ center of corn kernels. This determination was based on the requirement that the visible embryo is parallel to the camera’s plane.

In a typical downstream application, this visual information provided by image segmentation would be combined with data from the hyperspectral imaging sensor provided by QSorter, but with such a sensor having its field of view limited to only the middle pane. However, the other two panes still provide useful visual information for our classification models since the determination of the orientation of any particular kernel is not limited to only the frontal view of the kernel. Figure 2A shows oriented kernels, noting

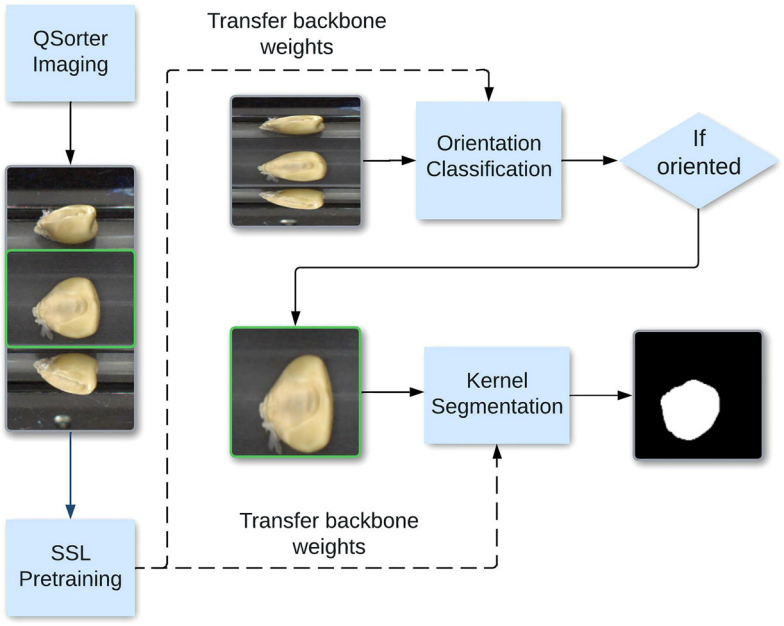

FIGURE 1
End-to-end pipeline for corn kernel classification and segmentation. The curved arrow shows the middle pane being processed for the segmentation task. We show that the classification and segmentation models perform bets with self-supervised weights.

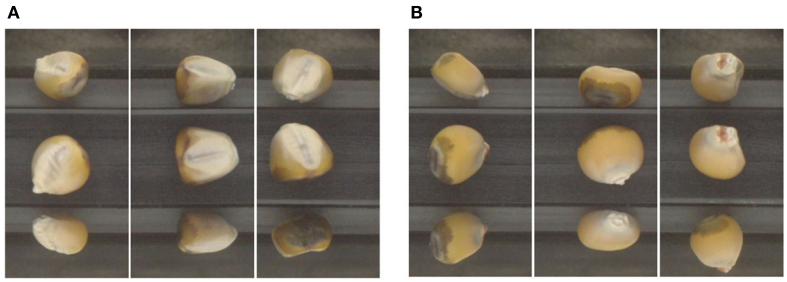

FIGURE 2
Images classified as "oriented" with the embryo visible **(A)** and "non-oriented" with the embryo not visible **(B)**.

the lighter portion visible in each middle pane, which is the corn embryo's visible part. Figure 2B shows non-oriented kernels in which the embryos are not visible or only partially visible.

2.1.2 Dataset for embryo segmentation

The embryo segmentation dataset consists of only 401 RGB images of corn kernels, taken from the same source of QSorter images as in the classification dataset above, along with their respective binary masks. Thus, the 2D image shapes were again 492 x 240. Segmentation (into the binary mask) distinguishes between the embryo and the rest of the background (including the non-embryo portion of the kernel). Figure 3 illustrates the segmentation annotation process for an RGB image and its mask. The three frames of each original (492, 240) dataset image were split into three individual images and downsampled to (128, 128). All completely negative masks and their respective RGB images were then removed.

2.2 SSL pretraining

2.2.1 Methods overview

The contrastive learning framework is a self-supervised learning method that maximizes the similarity between representations of an image and the augmented version of an image while minimizing the similarity between an image and other images (Zhao et al., 2021). The two models used for self-supervised pretraining were SimCLR (Simple Framework for Contrastive Learning of Visual Representations) (Chakraborty et al., 2020) and NNCLR (Nearest-Neighbor Contrastive Learning of Visual Representations) (Dwibedi et al., 2021). Figure S1 shows these two models superimposed on the same diagram.

SimCLR trains a backbone used for downstream processes by considering the contrastive loss of the representations of two distinct augmentations of images extracted from any given batch. If the initial images are the same, the pair of representations is considered a positive pair for the final calculation, and if the views are augmentations of two distinct images in the batch, then it is considered a negative pair. The representations are created by taking each augmented view of the initial image along a path including two networks: a base encoder where the desired backbone resides and a final projection head to calculate the contrastive loss of the representation in a separate space. NNCLR is also a contrastive model but differs from SimCLR in that upon taking both views of a given image through an encoder; the nearest neighbor algorithm is used to sample dataset representations for one of the views from a subset of the initial dataset. These are treated as the analog of the positive pairs described in the SimCLR model. Negative pairs are then the nearest neighbors of distinct initial images. Both architectures use the same InfoNCE loss to

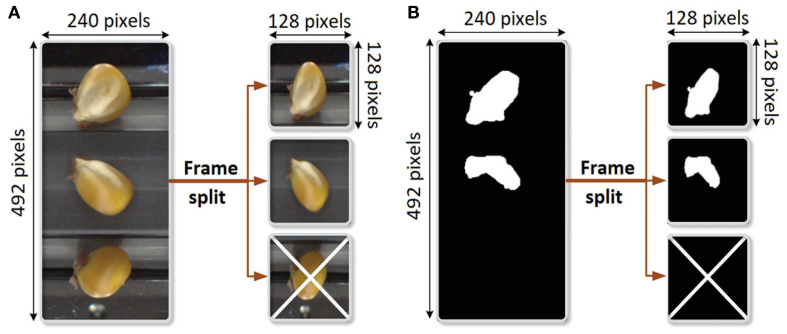

FIGURE 3
Preprocessing for segmentation consists of splitting each dataset image into three 128 x 128 images. Completely negative masks were excluded. **(A)** Preprocessing of RGB image. **(B)** Preprocessing of the mask.

maximize agreement, a loss function using categorical cross-entropy to maximize agreement with positive samples, commonly used in self-supervised learning (Song and Ermon, 2020). To evaluate the performance of the pretrained models, a linear probe — separate from the non-linear projection head included in both models — was attached directly to the encoder and was weight-updated at each step. The backbone and probe were then extracted to calculate validation accuracy for model selection.

2.2.2 Contrastive data augmentation

In many supervised image processing and computer vision tasks, data augmentation is used for the dual purposes of increasing the size of a labeled dataset through synthetic means and improving the diversity of a dataset. For purely supervised purposes, data augmentation can synthetically multiply the dataset's size by altering existing data and increasing the diversity of data to generalize the training set better (Wang and Perez, 2017). Contrastive learning uses heavier image augmentations than would normally be supplied to purely supervised training (Xie et al., 2020). This is due to the reliance of contrastive learning on using augmentations as a model for learning invariance to "style" changes, while the "content" component of a representation remains invariant (Doersch et al., 2015). Thus, heavy stylistic changes should generally benefit the learned representations.

The data augmentations used for our pretraining process were derived from the recommended augmentations particular to SimCLR, consisting of random zoom, random flip, color jitter, and Gaussian noise. NNCLR is less dependent in its performance than SimCLR on the precise type and magnitude of data augmentations used in training; indeed, upon applying augmentations to NNCLR pretraining similar to the full set recommended for SimCLR produced only a 1.6% performance improvement when compared to using only random crop (Dwibedi et al., 2021).

2.2.3 Pretraining setup

Hyperparameter sweeping during pretraining consisted of the variation of the contrastive learning rate, the type of weight initialization applied to the ResNet50 backbone, and data augmentation strength. The learning rate was chosen between 1e-3 and 1e-4, coupling the contrastive learning rate with the classification learning rate of the linear probe. Weight initialization was chosen between ImageNet and random initialization. The data augmentation strength of each augmentation was varied together and explained below. Thus, eight runs were processed for each sweep, and each sweep was repeated three times to ensure precision.

2.3 Classification

2.3.1 Data split

Of the 2697 images manually classified from the unlabeled dataset, there were 1300 oriented images and 1367 non-oriented images. Of the labeled images, 1697 were used for training, with an 800:897 class split in favor of non-oriented images. The rest were divided between validation and testing and were split evenly between the classes. So, 500 images were allocated to each set, with 250 images in each class. During pretraining, the images allocated to the validation and testing were separated from the unlabeled dataset used for contrastive learning, while the labeled training dataset was included, such that 43,286 out of the 44286 total images were used for unlabeled contrastive learning.

2.3.2 Training setup

The training process was set up to facilitate comparison between different models after undergoing end-to-end finetuning. Only ResNet50 was used for the backbones, as is standard in self-supervised model evaluation and as was used in both the NNCLR and SimCLR original papers (Chakraborty et al., 2020; Dwibedi et al., 2021; Shafiq and Gu, 2022). Two backbones for the end-to-end process were chosen from a pretraining sweep with the mentioned self-supervised contrastive architectures, and one backbone was initialized with ImageNet weights.

Data augmentation strength was defined separately for each particular augmentation depending on its configuration specifics: Random zoom acted by cropping to a single rectangle with its shape uniformly chosen between a maximum area of the initial 128x128 2D image shape and a minimum area of either 25% or 75% of the maximum area. Brightness and color transform was accomplished first by taking an identity matrix multiplied by the chosen brightness factor, then adding a matrix with uniformly chosen values selected between the jitter factor and its negative, and secondly by multiplying the original dataset image by this matrix. Brightness jitter increased the brightness of the image by either 50% or 75%, and the jitter factor was either 0.3 or 0.45. Gaussian noise was applied with a standard deviation of either 0.1 or 1.5. The only augmentation kept constant was random flip, constantly at 50% activation chance. Upon evaluation, the two chosen models from this pretraining sweep process—corresponding to the top-left-most two light-green boxes in Figure S2—were backbones pretrained by NNCLR with random initialization at LR = $1e^{-3}$ and SimCLR with ImageNet initialization at LR = $1e^{-3}$.

2.3.3 Feature extraction and finetuning

During training, separate trials were performed for each proportion of annotated data used in classification (1%, 10%, 25%, 100%). As in pretraining, each trial was repeated three times. With 1% and 10% data, a batch size of 4 was used; for 25% data, a batch size of 32 was used; and for 100% data, a batch size of 128 was used. During feature extraction, first, the ResNet-50 backbone from each initialization method was frozen to weight updates, upon which a trainable one-node classifier was constructed with sigmoid activation. Each classifier in every trial was trained for 300 epochs. In finetuning, the backbone was unfrozen, and the entire model was trained for 400 epochs. The same learning rate schedule was used in both phases at the fixed schedule of a 0.5 multiplier every 50 epochs. This process is illustrated in Figure 4.

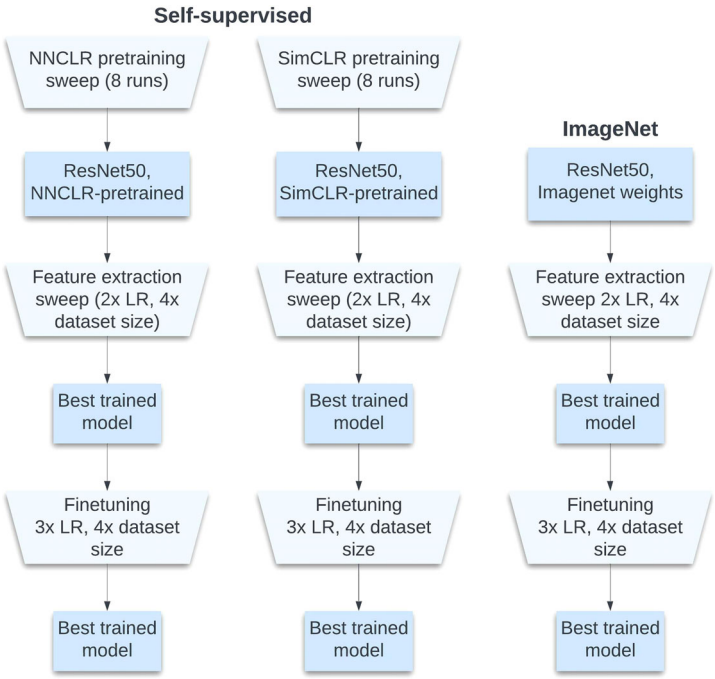

FIGURE 4
Training process. With each sweep over hyperparameters, the best model is chosen for the next round.

2.4 Segmentation

Semantic segmentation is a pixel-level classification problem where the goal is to assign a class label to each pixel of the image. Semantic segmentation of the classified images with the model created above is its natural downstream application. In doing so, full utilization of the QSorter pipeline can be achieved, where along with the immediate results of seed embryo pixel identification, these results can be combined with hyperspectral imaging data in a simple regression problem to pair results in segmentation with results in direct imaging.

2.4.1 Evaluation metrics

The Sørensen–Dice coefficient, also known as the Dice Similarity Coefficient (DSC), is a metric often used in segmentation tasks to evaluate the spatial overlap between two image masks (Taha and Hanbury, 2015). It is given by the equation below:

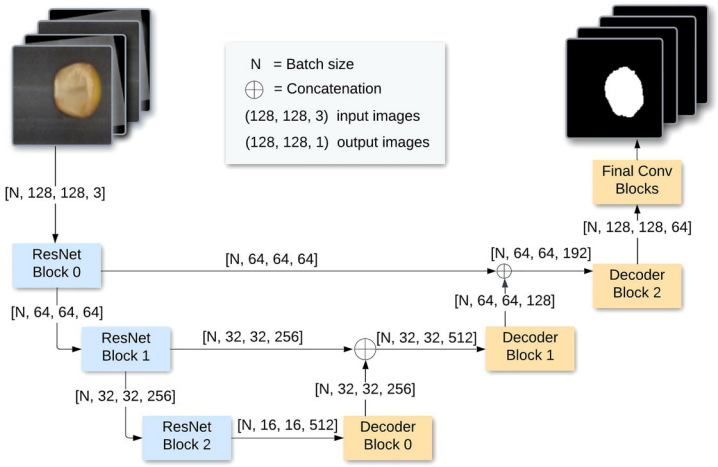

FIGURE 5
U-Net with ResNet50 backbone using filters of channel dimensions [64, 128, 256, 512].

$$Dice(\overrightarrow{y_1}, \overrightarrow{y_2}) = \frac{2\overrightarrow{y_1} \cdot \overrightarrow{y_2}}{\overrightarrow{y_1} + \overrightarrow{y_2}} \quad (1)$$

Here, $\overrightarrow{y_1}$ and $\overrightarrow{y_2}$ are the mask tensors flattened to one dimension. In statistical validation for computer vision tasks, DSC is often preferred over the pixel accuracy metric because DSC ignores true negatives, and pixel classes are often heavily biased toward the (negative) background, especially in binary semantic segmentation.

2.4.2 Model details

U-Net is a convolutional neural network commonly used for semantic segmentation tasks (Zunair and Hamza, 2021). It consists of a symmetric encoder-decoder pair, where the encoder down-samples while increasing the number of channels until a bottleneck tensor, from which the decoder up-samples while reducing the number of channels. For the segmentation task, we used U-Net with ResNet50 used as the encoder to both utilize and compare the self-supervised weights learned during the classification phase, as has been implemented in the literature to considerable advantage (Siddique et al., 2021). In this architecture, the encoder and decoder are not symmetric, as opposed to standard U-Net without a backbone, but skip connections are still fully implemented by limiting the depth of the encoder. Figure 5 shows a U-Net with a ResNet50 as its encoder and four sets of multi-channel feature maps.

2.4.3 Data augmentation

Data augmentation was applied to each training batch to increase the set of distinct training images and to reduce overfitting. Augmentations were coupled between any RGB image and its mask. All augmentations were executed with a 50% application chance. These consisted of combinations of the following: 1) horizontal flip across the vertical middle axis, 2) paired brightness and contrast transform with an application factor uniformly selected from [-0.2, 0.2], and 3) paired scaling and shearing affine transform, the scaling factor uniformly selected from [0.75, 1] and the shear angle uniformly selected from [-π/6, π/6]. Figure S3 shows an example of an augmented image-mask pair.

2.4.4 Training process

Due to the smaller size of the segmentation dataset compared to the classification dataset, ten-fold cross-validation was performed. Using ten folds, ten models were created separately for each backbone and each set of hyperparameters, repeated for each of the three weight initialization types, each trained on a train/validation split of 288/32. With every ten folds, the highest average Dice score across all ten was collected. A model with this set of best-performing hyperparameters was trained on all training data without a validation set for 300 epochs. This model was then evaluated on the full test set. Figure S4 illustrates the cross-validation process. Training and experiments were completed using Google Colab with NVIDIA Tesla T4 and K80 GPUs on 32 GB RAM.

3 Results and discussion

3.1 Classification results

3.1.1 Feature extraction evaluation

We first illustrate the impact of SSL pretraining on annotation efficiency, especially when compared with standard supervised approaches. Figure 6 compares the results of the classifier at various % of training data using a standard supervised loss vs both SimCLR and NNCLR. After feature extraction, (before end-to-end finetuning), both SimCLR and NNCLR were more annotation-efficient and performed better than purely transfer learning-based methods. Listing test results from greatest to least utilization of total available annotated data, the NNCLR-pretrained model had accuracies of 83.6%, 83.2%, 82.0%, and 76.2%; the SimCLR-pretrained model had accuracies of 83.0%, 82.6%, 81.4%, and 78.6%; and the ImageNet-initialized model had accuracies of 82.6%, 81.2%, 77.2%, and 74.6%. At every annotation percentage, the self-supervised models outperformed the ImageNet-based model, with the largest difference at 10% annotation, where the NNCLR-pretrained model outperformed the ImageNet-based model by 4.8%.

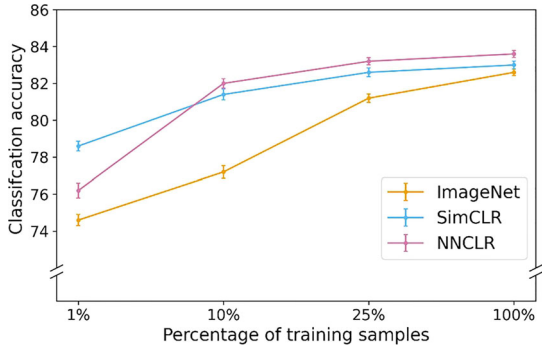

FIGURE 6
Feature extraction classification accuracy versus percentage of training samples for three types of weight initializations.

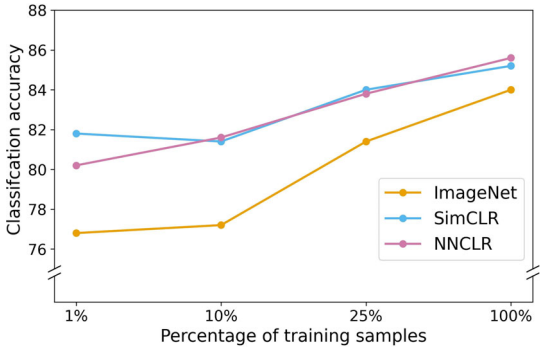

FIGURE 7
Classification accuracy versus percentage of training samples for three types of weight initializations.

3.1.2 Finetuning evaluation

After end-to-end finetuning, both SimCLR and NNCLR were more annotation-efficient and performed better than purely transfer learning-based methods, as shown in Figure 7. Listing test results from greatest to least utilization of total available annotated data, the NNCLR-pretrained model had accuracies of 85.6%, 83.8%, 81.6%, and 80.2%; the SimCLR-pretrained model had accuracies of 85.2%, 84.0%, 81.4%, and 81.8%; and the ImageNet-initialized model had accuracies of 84.0%, 81.4%, 77.2%, and 76.8%. At every annotation percentage, the self-supervised models outperformed all other models, with the largest difference at 1% annotation, where the SimCLR-pretrained model outperformed the ImageNet-based model by 5.0%. Furthermore, at just 1% annotation, SimCLR out-performs the ImageNet-initialized model at 25% annotation. At just 10% annotation, NNCLR also out-performs the ImageNet-initialized model at 25% annotation. We remind the reader that the total available annotated data is only around 5% of the total data (2697 annotated images out of 44,286 total images). SSL pretraining provides a significant boost in model performance, especially at very low total annotated data availability; for instance, a 10% usage of annotated data represents just 270 annotated images!

3.2 Comparisons

Models pretrained with contrastive SSL outperformed transfer learning models in every trial and between all data splits. Table 1 shows the performances of each model compared to the ImageNet-pretrained model. The results of the SimCLR and NNCLR pretrained models outperforming the transfer learning model and being more annotation efficient are clear. The performances of NNCLR and SimCLR were similar to each other among the four annotation percentages, but in training on the full dataset, NNCLR performed slightly better, while SimCLR was more efficient at the lowest data split.

3.3 Segmentation results

Figure 8 shows the test dataset evaluation results after the best models were selected and then finetuned, according to data from the previous three tables. It also shows the validation statistics and hyperparameter set for the chosen model. In Supplementary Information, Table S1 shows the averaged results from 10-fold cross-validation on U-Net with a ResNet50 backbone from weights pretrained with SimCLR, pretrained with NNCLR, and pretrained from ImageNet. Table S2 shows the selected models' hyperparameter set. The U-Net with a SimCLR-pretrained backbone trained at 1e-04 LR and four encoder-decoder filters performed best, with a test DICE score of 0.81 compared to an ImageNet-pretrained backbone at 0.78 DICE score.

The results from this section have a twofold implication: 1) they show U-Net with a backbone loaded with self-supervised pretrained weights can perform well, producing ~0.81 Dice score, and 2) they show semantic segmentation with these backbones outperform those pre-trained with ImageNet. Figure 9 displays three representative results from the segmentation model, including the predicted mask, the true mask, and the input RGB image.

TABLE 1 Relative performance by the accuracy of SimCLR-pretrained and NNCLR-pretrained models as compared to ImageNet preloaded model.

Pretraining	Percentage annotated data used			
	1%	10%	25%	100%
SimCLR	+3.4%	+4.2%	+2.6%	+1.2%
NNCLR	+5.0%	+4.4%	+2.4%	+1.4%

Each entry represents a performance gap in Figure 9. Available annotated dataset size is 2697.

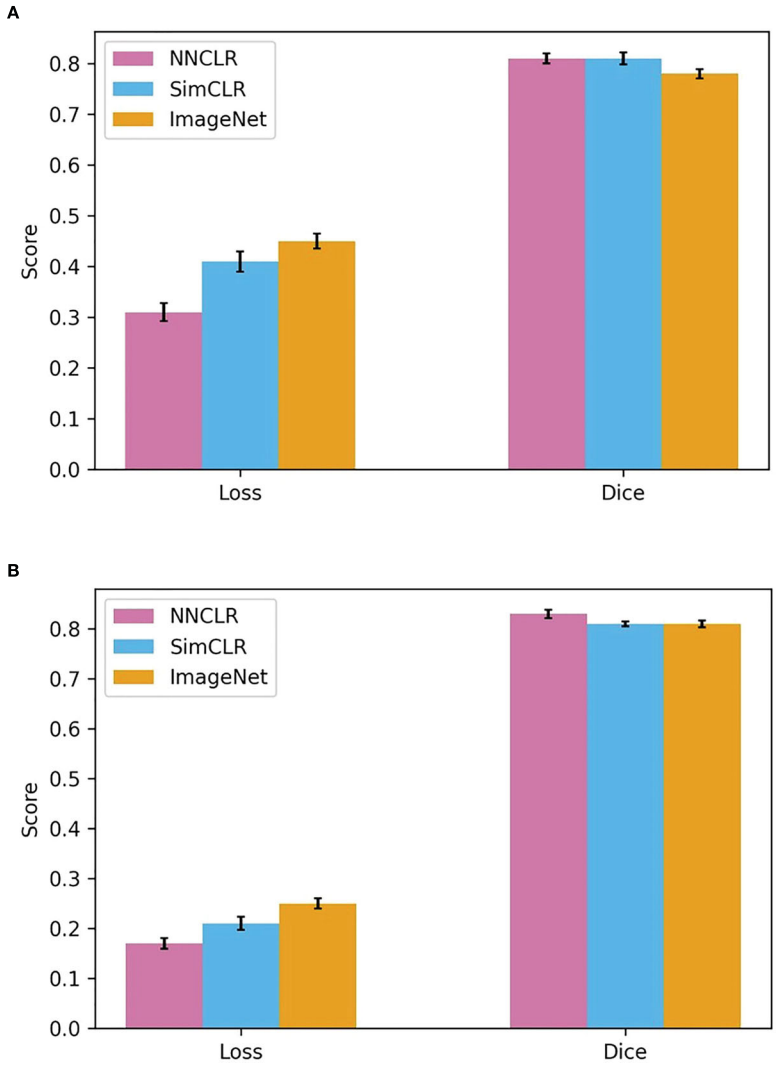

FIGURE 8
Loss and Dice scores for best hyperparameter sets for each weight initialization type. **(A)** Test set results, **(B)** Validation set results.

3.4 Advantages and limitations

In Section 3.1.1, we showed that a SimCLR-pretrained classifier that has gone through end-to-end finetuning out-performs an ImageNet-initialized classifier which uses 96% more annotated training data – the 1% annotation used by a SimCLR-pretrained model resulting in higher accuracy than the 25% annotation used by an ImageNet-initialized model. This is a clear example of the advantage of self-supervised contrastive methods in terms of both human-annotated data efficiency and accuracy. Not only does this curtail the time, labor, and resource-intensive process of annotation as described in the Introduction, but several other by-products of human annotation. For instance, label noise, data bias, the need for domain experts, and imperfect datasets in general are often inevitable with the use of large amounts of annotated data.

Other self-supervised methods have also been developed for computer vision tasks. Our experiments with non-contrastive methods such as SimSiam (Chen and He, 2021) turned out to be examples of the well-known faults of model collapse in non-contrastive self-supervised methods, with models consistently predicting uniform classes, reaching binary classification accuracies of no greater than 55%. We suggest that non-contrastive methods are particularly susceptible to collapse when applied to datasets with relatively homogenous feature spaces such as the applied corn kernel dataset. Furthermore, methods like inpainting (Pathak et al., 2016) have been shown to have poor performance in many applications compared to image augmentation-based methods. Thus, contrastive self-supervised methods which use pretext tasks similar to those of the strong augmentations we applied are particularly suited for processing plant datasets of little species or orientation variation.

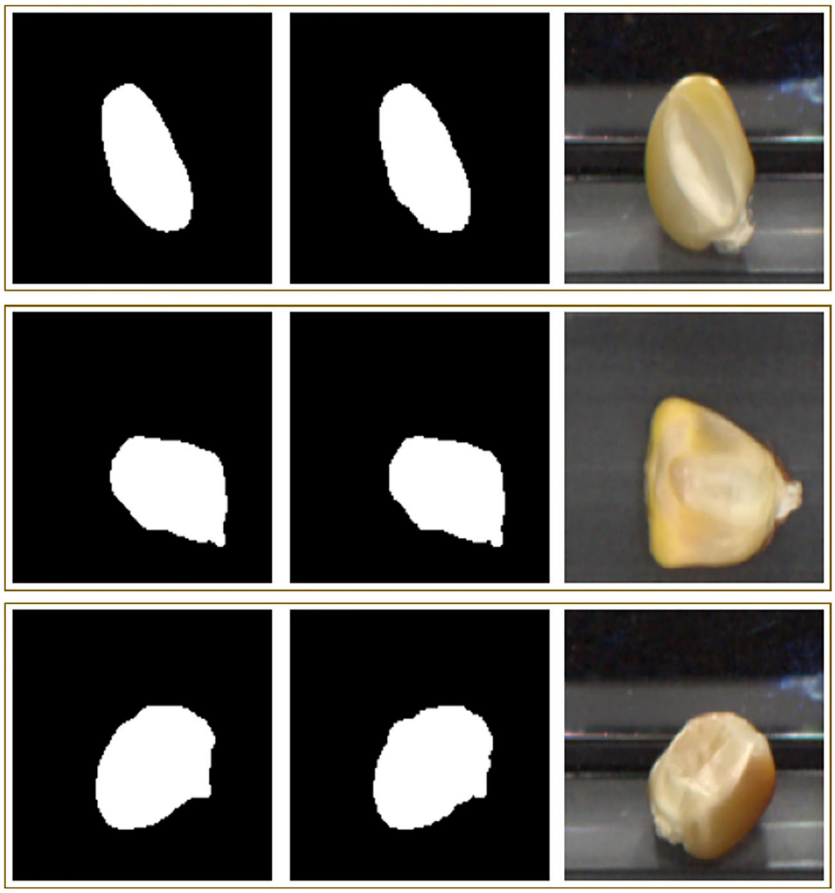

FIGURE 9
Three representative rows of segmentation inputs and outputs. The first column shows the predicted mask, the second shows the true mask, and the third shows the RGB input image.

Although we have found improved performance in applying self-supervised pretraining with all tasks, we expect monotone improvement in fine-tuned performance for classification and segmentation by increasing the size of unlabeled dataset. The clear advantage in relying on pretrained models is that procuring such data is far easier than with similar amounts of labeled data, as would be needed to improve purely supervised classification accuracy. Finally, we expect that such methods to be easily applied to, and very useful to a broad range of plant phenotyping applications. Recent examples of successful applications of such SSL training strategies include disease classification (Nagasubramanian et al., 2022) and insect detection (Kar et al., 2021).

4 Conclusion

From training contrastive learning models and comparing them with purely supervised and transfer learning methods, we found that self-supervised learning produces successful representations of an agricultural dataset applicable for downstream applications. We showed that NNCLR and SimCLR methods performed significantly better than their supervised counterparts, especially for the classification problem. These results also support the usage of strong augmentations in contrastive learning—far stronger than in end-to-end finetuning. In segmentation, self-supervised methods significantly improved over ImageNet pretraining, resulting in accurate masking capabilities and relative embryo size calculation. The combined results further show the transferable nature of self-supervised training. In particular, we illustrated that a single SSL-pretrained model (ResNet50 backbone) could be finetuned and used for two distinct downstream tasks – classification and segmentation. Furthermore, SSL pretraining allowed us to train models with very competitive performance even with very low amounts of total annotated data, for instance, with less than 1% (~400 out of 44000 total images) of annotation. Thus, we have demonstrated that self-supervised learning provides a meaningful path forward in advancing agricultural efficiency with computer vision and machine learning.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://zenodo.org/record/7577017, https://github.com/ddavidd23/ssl_corn.git.

Author contributions

UF, TJ, TL, and BG conceived the project. RW, UF, and TJ conducted physical experiments, data collection and data curation. RW and UF annotated the ground truth images. DD, KN, and TJ developed the machine learning framework. DD performed computational experiments, trained models, and analyzed results. DD wrote the manuscript draft with supervision from BG. All authors revised and edited the manuscript.

Funding

This work was partially supported by AI Institute for Resilient Agriculture (USDA-NIFA #2021-67021-35329), COALESCE: COntext Aware LEarning for Sustainable CybEr-Agricultural Systems (CPS Frontier # 1954556). BG and TL acknowledge support from PSI faculty fellowship.

Acknowledgments

We thank Dr. Candice Gardner at the USDA-ARS Plant Introduction Station.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1108355/full#supplementary-material

References

Aboobucker, S. I., Jubery, T. Z., Frei, U. K., Chen, Y. R., Foster, T., Ganapathysubramanian, B., et al. (2022). "Protocols for in vivo doubled haploid (DH) technology in maize breeding: From haploid inducer development to haploid genome doubling," in *Plant gametogenesis* (Humana, New York, NY: Clifton, NJ), 213–235.

Armstrong, P. R., Tallada, J. G., Hurburgh, C., Hildebrand, D. F., and Specht, J. E. (2011). Development of single-seed near-infrared spectroscopic predictions of corn and soybean constituents using bulk reference values and mean spectra. *Trans. ASABE* 54, 1529–1535. doi: 10.13031/2013.39012

Boote, B. W., Freppon, D. J., de la Fuente, G. N., Lübberstedt, T., Nikolau, B. J., and Smith, E. A. (2016). Haploid differentiation in maize kernels based on fluorescence imaging. *Plant Breed.* 135, 439–445. doi: 10.1111/pbr.12382

Chaikam, V., Molenaar, W., Melchinger, A. E., and Boddupalli, P. M. (2019). Doubled haploid technology for line development in maize: technical advances and prospects. *Theor. Appl. Genet.* 132, 3227–3243. doi: 10.1007/s00122-019-03433-x

Chakraborty, S., Gosthipaty, A. R., and Paul, S. (2020). "G-SimCLR: Self-supervised contrastive learning with guided projection *via* pseudo labelling," in *2020 International Conference on Data Mining Workshops (ICDMW)*. 912–916 (Sorrento, Italy: IEEE).

Chen, X., and He, K. (2021). "Exploring simple Siamese representation learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15745–15753.

Dobrescu, A., Valerio Giuffrida, M., and Tsaftaris, S. A. (2019). "Understanding deep neural networks for regression in leaf counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Long Beach, CA, USA. 4321–4329.

Doersch, C., Gupta, A., and Efros, A. A. (2015). "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE International Conference on Computer Vision*. 1422–1430.

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., and Zisserman, A. (2021). "With a little help from my friends: Nearest-neighbor contrastive learning of visual representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9588–9597.

Fahlgren, N., Gehan, M. A., and Baxter, I. (2015). Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. *Curr. Opin. Plant Biol.* 24, 93–99. doi: 10.1016/j.pbi.2015.02.006

Ghosal, S., Zheng, B., Chapman, S. C., Potgieter, A. B., Jordan, D. R., Wang, X., et al. (2019). A weakly supervised deep learning framework for sorghum head detection and counting. *Plant Phenomics* 2019, 1525874. doi: 10.34133/2019/1525874

Güldenring, R., and Nalpantidis, L. (2021). Self-supervised contrastive learning on agricultural images. *Comput. Electron. Agric.* 191, 106510. doi: 10.1016/j.compag.2021.106510

Guo, W., Carroll, M. E., Singh, A., Swetnam, T. L., Merchant, N., Sarkar, S., et al. (2021). UAS-based plant phenotyping for research and breeding applications. *Plant Phenomics*, 2021:9840192. doi: 10.34133/2021/9840192

Gustin, J. L., Frei, U. K., Baier, J., Armstrong, P., Lübberstedt, T., and Settles, A. M. (2020). Classification approaches for sorting maize (Zea mays subsp. mays) haploids using single-kernel near-infrared spectroscopy. *Plant Breed.* 139, 1103–1112. doi: 10.1111/pbr.12857

Halcro, K., McNabb, K., Lockinger, A., Socquet-Juglard, D., Bett, K. E., and Noble, S. D. (2020). The BELT and phenoSEED platforms: shape and colour phenotyping of seed samples. *Plant Methods* 16, 1–13. doi: 10.1186/s13007-020-00591-8

Jiang, Y., and Li, C. (2020). Convolutional neural networks for image-based high-throughput plant phenotyping: a review. *Plant Phenomics* 2020, 4152816. doi: 10.34133/2020/4152816

Jiang, H. Y., Zhu, Y. J., Wei, L. M., Dai, J. R., Song, T. M., Yan, Y. L., et al. (2007). Analysis of protein, starch and oil content of single intact kernels by near infrared reflectance spectroscopy (NIRS) in maize (Zea mays l.). *Plant Breed.* 126, 492–497. doi: 10.1111/j.1439-0523.2007.01338.x

Jing, L., and Tian, Y. (2020). "Self-supervised visual feature learning with deep neural networks: A survey," in *IEEE transactions on pattern analysis and machine intelligence*, Vol. 43. 4037–4058.

Jones, R. W., Reinot, T., Frei, U. K., Tseng, Y., Lübberstedt, T., and McClelland, J. F. (2012). Selection of haploid maize kernels from hybrid kernels for plant breeding using near-infrared spectroscopy and SIMCA analysis. *Appl. Spectrosc.* 66, 447–450. doi: 10.1366/11-06426

Jubery, T. Z., Carley, C. N., Singh, A., Sarkar, S., Ganapathysubramanian, B., and Singh, A. K. (2021). Using machine learning to develop a fully automated soybean nodule acquisition pipeline (snap). *Plant Phenomics* 2021, 9834746. doi: 10.34133/2021/9834746

Kar, S., Nagasubramanian, K., Elango, D., Nair, A., Mueller, D. S., O'Neal, M. E., et al. (2021). "November. self-supervised learning improves agricultural pest classification," in *AI for Agriculture and Food Systems*.

Kattenborn, T., Leitloff, J., Schiefer, F., and Hinz, S. (2021). Review on convolutional neural networks (CNN) in vegetation remote sensing. *ISPRS J. Photogrammetry Remote Sens.* 173, 24–49. doi: 10.1016/j.isprsjprs.2020.12.010

Körschens, M., Bodesheim, P., Römermann, C., Bucher, S. F., Migliavacca, M., Ulrich, J., et al. (2021). "Weakly supervised segmentation pretraining for plant cover prediction," in *DAGM German Conference on Pattern Recognition*, Cham. 589–603 (Bonn, Germany: Springer).

Lee, U., Kwon, H., Wu, M., and Wang, M. (2021). Retrospective analysis of the US corn ethanol industry for 2005–2019: implications for greenhouse gas emission reductions. *Biofuels bioprod. Biorefining* 15, 1318–1331. doi: 10.1002/bbb.2225

Lin, X., Li, C. T., Adams, S., Kouzani, A. Z., Jiang, R., He, L., et al. (2023). Self-supervised leaf segmentation under complex lighting conditions. *Pattern Recognition* 135, 109021. doi: 10.1016/j.patcog.2022.109021

Liu, Q., Wang, Z., Long, Y., Zhang, C., Fan, S., and Huang, W. (2022). Variety classification of coated maize seeds based on raman hyperspectral imaging. *Spectrochim. Acta - A: Mol. Biomol. Spectrosc.* 270, 120772. doi: 10.1016/j.saa.2021.120772

Marin Zapata, P. A., Roth, S., Schmutzler, D., Wolf, T., Manesso, E., and Clevert, D. A. (2021). Self-supervised feature extraction from image time series in plant phenotyping using triplet networks. *Bioinformatics* 37, 861–867. doi: 10.1093/bioinformatics/btaa905

McClure, W. F. (2003). 204 years of near infrared technology: 1800–2003. *J. Near Infrared Spec.* 11, 487–518. doi: 10.1255/jnirs.399

Melchinger, A. E., Munder, S., Mauch, F. J., Mirdita, V., Böhm, J., and Mueller, J. (2017). High-throughput platform for automated sorting and selection of single seeds based on time-domain nuclear magnetic resonance (TD-NMR) measurement of oil content. *Biosyst. Eng.* 164, 213–220. doi: 10.1016/j.biosystemseng.2017.10.011

Misra, I., and Maaten, L. V. D. (2020). "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6707–6717.

Mohanty, S. P., Hughes, D. P., and Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Front. Plant Sci.* 7, 419. doi: 10.3389/fpls.2016.01419

Nagasubramanian, K., Jubery, T., Fotouhi Ardakani, F., Mirnezami, S. V., Singh, A. K., Singh, A., et al. (2021). How useful is active learning for image-based plant phenotyping? *Plant Phenome J.* 4, e20020. doi: 10.1002/ppj2.20020

Nagasubramanian, K., Singh, A., Singh, A., Sarkar, S., and Ganapathysubramanian, B. (2022). Plant phenotyping with limited annotation: Doing more with less. *Plant Phenome J.* 5, e20051. doi: 10.1002/ppj2.20051

Naik, H. S., Zhang, J., Lofquist, A., Assefa, T., Sarkar, S., Ackerman, D., et al. (2017). A real-time phenotyping framework using machine learning for plant stress severity rating in soybean. *Plant Methods* 13, 1–12. doi: 10.1186/s13007-017-0173-7

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., and Efros, A. A. (2016). "Context encoders: Feature learning by inpainting," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA 2536–2544.

Paulsen, M. R., and Hill, L. D. (1985). Corn quality factors affecting dry milling performance. *J. Agric. Eng. Res.* 31, 255–263. doi: 10.1016/0021-8634(85)90092-7

Pound, M. P., Atkinson, J. A., Wells, D. M., Pridmore, T. P., and French, A. P. (2017). "Deep learning for multi-task plant phenotyping," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*. Venice, Italy 2055–2063.

QualySense *QualySense*. Available at: https://qualysense.com/.

Satake-USA *Optical sorting & processing - rice, wheat & grains - satake USA*. Available at: https://satake-usa.com/ (Accessed 22 Nov. 2022).

Shafiq, M., and Gu, Z. (2022). Deep residual learning for image recognition: A survey. *Appl. Sci.* 12, 8972. doi: 10.3390/app12188972

Siddique, N., Paheding, S., Elkin, C. P., and Devabhaktuni, V. (2021). U-Net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access* 9, 82031–82057. doi: 10.1109/ACCESS.2021.3086020

Silvela, L., Rodgers, R., Barrera, A., and Alexander, D. E. (1989). Effect of selection intensity and population size on percent oil in maize, zea mays l. *Theor. Appl. Genet.* 78, 298–304. doi: 10.1007/BF00288815

Singh, A. K., Ganapathysubramanian, B., Sarkar, S., and Singh, A. (2018). Deep learning for plant stress phenotyping: trends and future perspectives. *Trends Plant Sci.* 23, 883–898. doi: 10.1016/j.tplants.2018.07.004

Song, J., and Ermon, S. (2020). Multi-label contrastive predictive coding. *Adv. Neural Inf. Process. Syst.* 33, 8161–8173.

Spielbauer, G., Armstrong, P., Baier, J. W., Allen, W. B., Richardson, K., Shen, B., et al. (2009). High-throughput near-infrared reflectance spectroscopy for predicting quantitative and qualitative composition phenotypes of individual maize kernels. *Cereal Chem.* 86, 556–564. doi: 10.1094/CCHEM-86-5-0556

Taha, A. A., and Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* 15, 1–28. doi: 10.1186/s12880-015-0068-x

Ubbens, J. R., and Stavness, I. (2017). Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks. *Front. Plant Sci.* 8, 1190. doi: 10.3389/fpls.2017.01190

Wang, J., and Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit.* 11, 1–8.

Wang, A., Zhang, W., and Wei, X. (2019). A review on weed detection using ground-based machine vision and image processing techniques. *Comput. Electron. Agric.* 158, 226–240. doi: 10.1016/j.compag.2019.02.005

Ward, O. P., and Singh, A. (2002). Bioethanol technology: developments and perspectives. *Adv. Appl. Microbiol.* 51, 53–80. doi: 10.1016/S0065-2164(02)51001-7

Weinstock, B. A., Janni, J., Hagen, L., and Wright, S. (2006). Prediction of oil and oleic acid concentrations in individual corn (Zea mays l.) kernels using near-infrared reflectance hyperspectral imaging and multivariate analysis. *Appl. Spectrosc.* 60, 9–16. doi: 10.1366/000370206775382631

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. (2020). Unsupervised data augmentation for consistency training. *Adv. Neural Inf. Process. Syst.* 33, 6256–6268.

Xu, Z., Fan, S., Liu, J., Liu, B., Tao, L., Wu, J., et al. (2019). A calibration transfer optimized single kernel near-infrared spectroscopic method. *Spectrochim. Acta - A: Mol. Biomol. Spectrosc.* 220, 117098. doi: 10.1016/j.saa.2019.05.003

Yang, G., Wang, Q., Liu, C., Wang, X., Fan, S., and Huang, W. (2018). Rapid and visual detection of the main chemical compositions in maize seeds based on raman hyperspectral imaging. *Spectrochim. Acta - A: Mol. Biomol. Spectrosc.* 200, 186–194. doi: 10.1016/j.saa.2018.04.026

Zhai, X., Oliver, A., Kolesnikov, A., and Beyer, L. (2019). "S4l: Self-supervised semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul, Korea (South) 1476–1485.

Zhao, X., Vemulapalli, R., Mansfield, P. A., Gong, B., Green, B., Shapira, L., et al. (2021). "Contrastive learning for label efficient semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10623–10633.

Zunair, H., and Hamza, A. B. (2021). Sharp U-net: depthwise convolutional network for biomedical image segmentation. *Comput. Biol. Med.* 136, 104699. doi: 10.1016/j.compbiomed.2021.104699