

# Cherry growth modeling based on Prior Distance Embedding contrastive learning: Pre-training, anomaly detection, semantic segmentation, and temporal modeling

Wei Xu <sup>a,b,1</sup>, Ruiya Guo <sup>b,1</sup>, Pengyu Chen <sup>b</sup>, Li Li <sup>c,b</sup>, Maomao Gu <sup>b</sup>, Hao Sun <sup>b</sup>, Lingyan Hu <sup>b,\*</sup>, Zumin Wang <sup>b,\*</sup>, Kefeng Li <sup>a,\*</sup>

<sup>a</sup> Faculty of Applied Sciences, Macao Polytechnic University, Macao Special Administrative Region of China

<sup>b</sup> School of Information Engineering, Dalian University, Dalian, China

<sup>c</sup> Shanghai Chenrui Communication Technology Company Ltd., Shanghai, China

## ARTICLE INFO

### Keywords:

Contrastive learning  
Prior Distance Embedding (PDE)  
Plant phenotyping  
Deep learning  
Plant temporal modeling

## ABSTRACT

In current plant phenotyping research, the study of plant time-series images based on deep learning has received widespread attention. While such image data is relatively easy to obtain, the cost of annotation is high. One efficient method for achieving cost-effective training is through contrastive learning. Plant growth is slow, and the changes in image sequences over a period of time are small, with simple semantic information. Previous contrastive pre-training models struggled to effectively distinguish positive samples from the same image with different augmented views and similar negative samples from different images. Therefore, this paper proposes a method called self-supervised contrastive learning method for plant time-series images with a Prior Distance Embedding (PDE). The semantic information in images corresponding to different phenological stages of plants varies. This method transforms this crucial domain knowledge into prior distances for image pairs and conducts contrastive learning pre-training. The learned weights can be transferred to downstream tasks. Building upon this method, experiments were conducted on cherry time-series images to assess the quality of pre-training through a plant phenotyping image semantic segmentation task. To provide a comprehensive example of plant time-series image phenotypic analysis, this paper establishes a cherry growth temporal model, specifically including PDE pre-training, anomaly detection, semantic segmentation, and recording the results from the temporal dimension. The experiments indicate that this self-supervised contrastive learning method can be effectively applied to the pre-training of plant time-series images, demonstrating broad applicability in various computer vision studies related to plant phenotyping.

## 1. Introduction

The observable external characteristics, as well as the internal physiological and biochemical features, of an organism, determined by its genotype and the environment in which it exists, are referred to as the plant phenotype (Houle et al., 2010). Understanding the phenotypic features and traits of plants is a crucial proposition in biological research. Without comprehensive phenotypic data, a profound understanding of the complex interplay (Houle et al., 2010; Pan et al., 2015) between the genome and environmental factors in shaping plant phenotypes would be unattainable. Traditional plant phenotypic analysis primarily involves manual measurement of various parameters, resulting in small-scale analysis, low efficiency, and significant potential for

errors. With the development of imaging sensors, computer vision-based methods for plant phenotypic analysis have rapidly increased. These methods are applied in various aspects of plant research (Kolhar and Jagtap, 2021), particularly in the study of specialty crops with high added value.

In current plant phenotyping research based on computer vision technology, there is widespread attention (Kierdorf et al., 2023; Sun et al., 2022) given to the study of growth modeling using time-series images due to the periodicity and rhythmicity of plant growth. The general research process typically involves two main stages: information extraction and temporal modeling. In the information extraction stage, digital image processing methods are commonly employed, especially image classification (Cao et al., 2021; Cui et al., 2021), object

\* Corresponding authors.

E-mail addresses: [hulingyan@dlu.edu.cn](mailto:hulingyan@dlu.edu.cn) (L. Hu), [wangzumin@dlu.edu.cn](mailto:wangzumin@dlu.edu.cn) (Z. Wang), [kefengl@mpu.edu.mo](mailto:kefengl@mpu.edu.mo) (K. Li).

<sup>1</sup> Equal Contribution.

detection (Cao and Xin, 2021), semantic segmentation (Song et al., 2022; Yasrab et al., 2021) and various deep learning techniques to extract phenotypic data from individual image data. In the temporal modeling stage, information is accumulated over time, merging data from different growth stages to establish specific models. This allows for joint analysis (Song et al., 2022; Sun et al., 2022) with other external factors.

During the collection of plant time-series images, devices typically automatically capture images of a specific area of the plant at fixed intervals. This process is convenient (Richardson et al., 2018). However, in the process of building deep learning models for extracting plant information, a significant number of labeled images are typically required for training. Due to the presence of numerous details in plant images, such as the edges of flowers and leaves, manually annotating datasets incurs high costs. This necessitates models that can achieve better training results (Güldenring and Nalpantidis, 2021) with fewer manually labeled data.

For the aforementioned situation, contrastive learning with unlabeled data for pre-training is one approach (Güldenring and Nalpantidis, 2021) to achieve label-efficient training. Currently, prominent contrastive learning models include SimCLR (Chen et al., 2020b), MoCo (He et al., 2020), SimSia (Chen and He, 2021) and so on. These models learn representations (Güldenring and Nalpantidis, 2021) by maximizing the consistency between different augmented views of the same data instance, and they have demonstrated favorable results when transferred to downstream tasks. Exploratory research on contrastive learning has already begun in the field of plant and crop images, such as in plant phenotypic segmentation (Güldenring and Nalpantidis, 2021), plant remote sensing (Güldenring and Nalpantidis, 2021), disease and pest monitoring (Kar et al., 2022), seed classification (Margapuri and Neilsen, 2021) and so on.

However, the number of studies on contrastive learning in the field of plants is far less than in other domains, and there is a lack of research on contrastive learning specifically applied to plant time-series images. Plant image sequences have their own specificity compared to regular images. Plant growth is slow, and the image sequences over a period of time exhibit minimal changes, making them relatively similar. Organs such as flowers, leaves, and trunks dominate the images, and the semantic information is straightforward. In conventional contrastive learning models, positive image pairs are formed by different data augmentation views from the same image, while negative image pairs consist of views from different images. The training objective is to bring the distance between positive sample pairs closer and push the distance between negative sample pairs further apart (Chen et al., 2020b; Chen and He, 2021; Grill et al., 2020; Güldenring and Nalpantidis, 2021; He et al., 2020). If plant sequence images are input into conventional contrastive learning models, due to the semantic similarity of some images, the model may struggle to determine whether a positive sample pair consists of different augmentations of the same image or if it is a negative sample pair from different yet similar images. This difficulty can hinder the convergence of the model, as observed in experiments. Conventional generic contrastive learning methods may not be suitable for pre-training on plant time-series images.

To enhance pre-training on plant time-series images, this paper proposes a method called self-supervised contrastive learning method for plant time-series images with a priori distance embedding (PDE). During the process of plant growth, there is a relatively fixed annual growth cycle, known as the phenological period. This period includes stages such as leafing, flowering, fruiting, and so on (Yasrab et al., 2021). The phenological period is essential domain knowledge in the field of botany. Embedding domain knowledge as priors into machine learning models can eliminate barriers between knowledge and data, providing the model with richer and more pattern-compliant information (Chen and Zhang, 2022). Different phenological stages exhibit distinct plant phenotypes, reflected in images through variations in the color, state, size, and proportion of different plant organs. This implies

a correlation between phenological stages and the semantic content in corresponding images. The method proposed in this paper extracts the phenological stages of plants (Yasrab et al., 2021) to establish prior knowledge with a hierarchical distance structure for image pairs. Subsequently, contrastive training is performed within the Siamese network (Bromley et al., 1993). The proposed approach introduces hierarchical distance and classification distance metrics to compute contrastive loss, aiming to converge the model towards the prior distance as the target. The trained encoder can effectively extract semantic information from plant images and be transferred to downstream tasks such as plant information extraction and time-series modeling.

Based on this approach, this paper provides an example of a full-process plant phenotyping analysis on the task of phenotypic time-series phenotyping of cherry. First, a surveillance camera was used and an acquisition system was built to obtain cherry time-series images. Second, self-supervised pre-training based on PDE was performed to obtain an encoder that can extract semantic information about cherries. Third, cherry growth modeling was started, and the pre-trained encoder was used as a feature extractor to detect anomalies in the plant temporal images and exclude the anomalous images. Fourth, the pre-trained encoder is migrated to the U-net network, and the semantic segmentation model is supervised to train the semantic segmentation model to segment the flowers, fruits, leaves, and trunks in the images. Fifth, temporal modeling is performed, and for each cherry image in the time series, the changes in the proportion of different organ regions and color are recorded in the time dimension, and Richards curves are fitted to establish a cherry growth model.

The remainder of this article is structured as follows. The second section (Section 2) provides an overview of existing self-supervised contrast learning methods with respect to plant phenotyping studies on time series images. The third section details the pre-training of comparative learning of plant time-series images embedded with a priori distance metric, including the overall training process, image pair construction and comparison loss function. Part IV performs cherry temporal sequence modeling based on PDE. Part V conducts ablation and comparison experiments on PDE and reports the results of PDE-based time-series modeling. Section 6 discusses the PDE algorithm and the timing modeling results based on the algorithm. Finally, Part VII summarizes the full paper.

## 2. Related work

### 2.1. Self-supervised contrastive learning

Self-supervised training methods based on noise-contrastive estimation are currently a significant research focus (Chen and He, 2021; He et al., 2020) in the field of computer vision. SimCLR (Chen et al., 2020b) has constructed a simple end-to-end visual representation contrastive learning framework, using the NT-Xent loss as the loss function. It employs a large batch size 4096 with augmented images as positive samples and the remaining images in the mini-batch as negative samples. MoCo (Chen et al., 2020a; He et al., 2020) treat contrastive learning as a dictionary query task. They introduce a queue to store the results of previous mini-batch training and use a momentum encoder to update the queue. The InfoNCE loss is calculated on the queue, and models can be trained with more common batch sizes 128–256. SimCLR and MoCo both require contrasting positive and negative samples during training to calculate the loss. However, subsequent approaches like BYOL (Grill et al., 2020) and SimSiam (Chen and He, 2021) demonstrated the feasibility of training solely with positive samples.

Exploratory research on contrastive learning in plant images has already begun. Güldenring and Nalpantidis (2021) employed the SwAV method (Caron et al., 2020) for self-supervised pre-training on four different types of agricultural image datasets. Kar et al. (2022) employed minimal labeling to classify 12 agricultural pests. They inputted

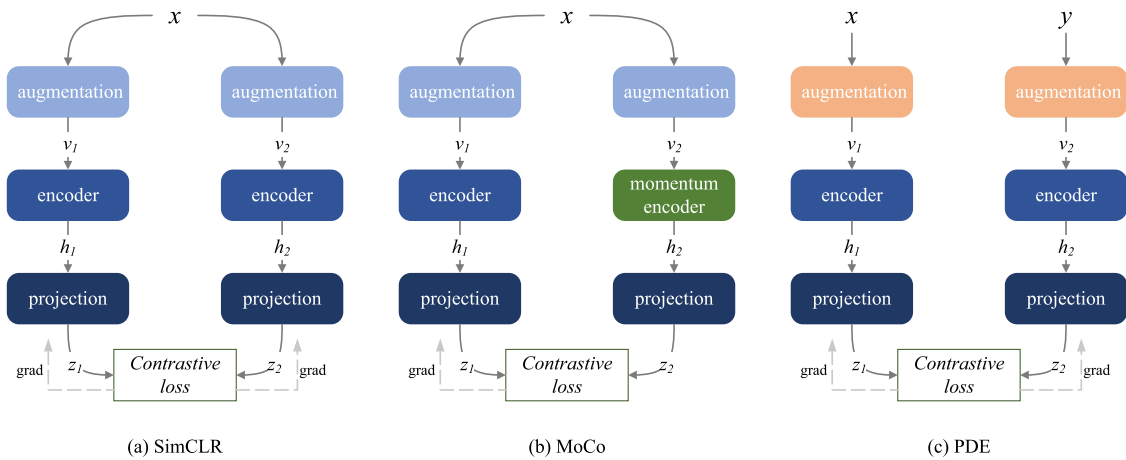


Fig. 1. Contrastive learning methods.

both the original images and images segmented using a locally entropy-based approach into the BYOL (Grill et al., 2020) method. Margapuri and Neilsen (2021) compared the effectiveness of SimCLR (Chen et al., 2020b), MoCo (He et al., 2020) and BYOL (Grill et al., 2020) in a seed classification task. They utilized all images in the dataset for self-supervised contrastive pre-training and fine-tuned the models with 5% of labeled data.

Currently, mainstream contrastive learning methods focus on contrasting views of a single image after data augmentation. These methods do not explicitly consider the scenario where there is a high degree of similarity among training samples, and there is a lack of specific contrastive learning research designed for the characteristics of plant images.

## 2.2. Plant phenotyping based on time-series images

In current plant phenotyping research based on computer vision technology, the use of time-series images for growth modeling has gained significant attention and application in recent years. One notable example is the phenological research based on the PhenoCam network, which includes 1783 observation sites. Every 30 min, digital cameras with fixed angles automatically capture plant images and upload them (Richardson et al., 2018; Seyednasrollah et al., 2019). Cui et al. (2021) proposed a forest phenology recognition method with strong fine-grained feature recognition capabilities. The study utilized images from oak and maple forests in PhenoCam as research materials. A deep learning classification model based on ResNet50 was designed, incorporating an attention mechanism to address the subtle differences in forest phenology. The approach achieved high accuracy in recognizing the phenological stages of trees. Song et al. (2022) explored the improvement of tropical phenology monitoring by employing various deep learning superpixel segmentation models to automatically distinguish leaf and non-leaf regions in Phenocam images. This approach allows for the derivation of the leaf proportion within the tree crown, leading to the temporal variation of Green Chromatic Coordinates Cao and Xin (2021) utilized the YOLOv3 model to identify the phenology of deciduous broadleaf forest vegetation in Phenocam images. This method enables the accurate and rapid localization of deciduous broadleaf forest areas, facilitating the extraction of phenological information.

Additionally, Kierdorf et al. (2023) introduced an open dataset called GrowliFlower, which consists of geo-referenced drone images of cauliflower phenotypes. The dataset includes ortho-RGB and multispectral images covering the entire growth period from planting to harvest in 2020 and 2021. Instance segmentation results based on Mask R-CNN are also provided. Aksoy et al. (2015) used a depth camera mounted on a robotic arm to capture infrared images of tobacco growth sequences.

They performed segmentation between different leaves and established a growth model by fitting ellipses to the segmented leaves. Yasrab et al. (2021) employed a generative adversarial network GAN to train a deep learning network using time-series images of plant growth. The network was used to predict and segment the appearance of future leaves and root systems.

In time-series growth modeling research, information extraction models based on deep learning are typically either trained from scratch without pre-training or undergo transfer learning using ImageNet weights. In contrast to the aforementioned approaches, we employ contrastive learning to obtain pre-trained weights specifically tailored for crops. Unsupervised learning finds particularly broad applications in smart agriculture since collecting images is relatively easy compared to the substantial effort required for manual annotation.

## 3. Self-supervised contrastive learning method for plant time-series images with a priori distance embedding

This study proposes a method called self-supervised contrastive learning method for plant time-series images with a Prior Distance Embedding. This approach transfers the prior domain knowledge of phenological stages into representation learning. This section presents the pre-training method in a top-down manner. Firstly, the entire contrastive learning framework is introduced, followed by detailed explanations of the key components, namely, image pair construction and the contrastive loss based on prior distance.

### 3.1. Contrastive learning framework

First, the phenological periods are determined by reading the plant time-series images. Four types of image pairs are generated based on this a priori information: the same period of the same sequence, different periods of the same sequence, different sequences of the same period, and different sequences of different periods. The process of generating image pairs from phenological periods is the incorporation of prior phenological data-contained domain knowledge into deep learning. In addition, the prior phenological information is recorded using various types of image pairs with varying priori distances. In Section 3.2 of this paper, the generation of contrastive learning image pairs is described in detail. Fig. 1(a) shows the general framework of PDE.

The contrastive model is a Siamese network (Bromley et al., 1993) with  $x$  and  $y$  as input images. After model input, data augmentation is necessary. In contrast to previous studies, the data augmentation in PDE serves to reduce the effect of color on model training, particularly the green color of large leaves, so that the model can concentrate on

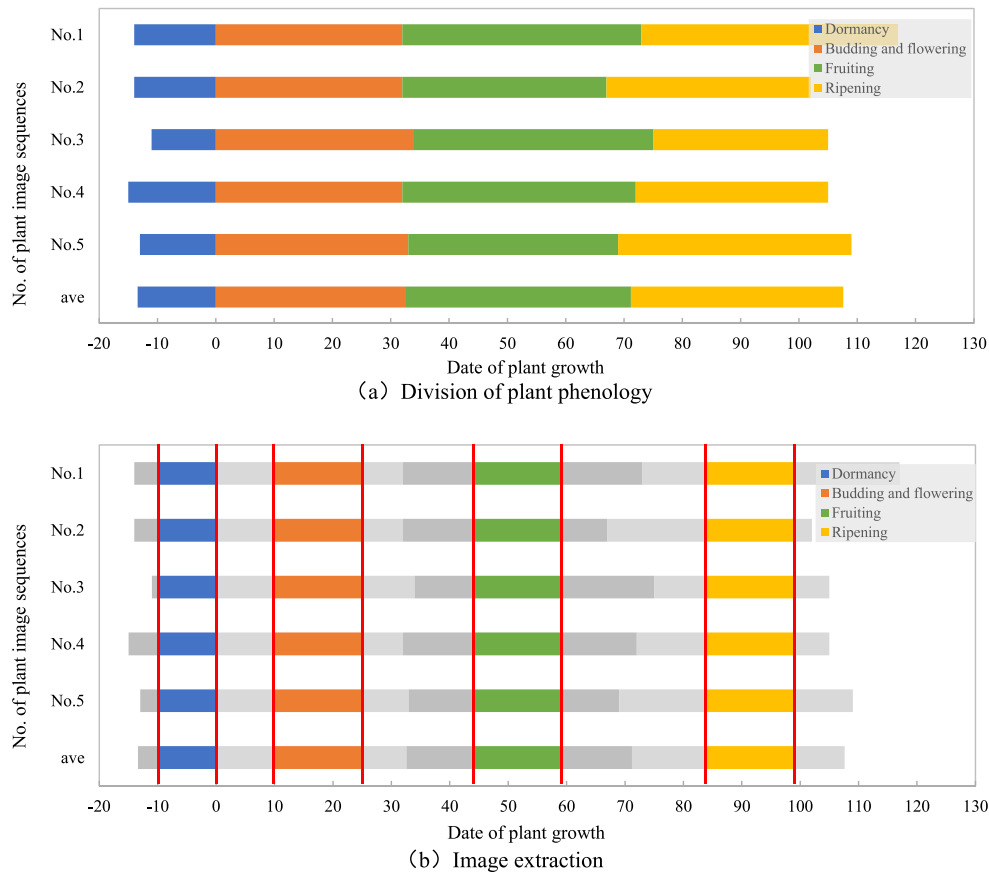


Fig. 2. Phenological stage division and image extraction.

higher-level semantic information other than color. The images after data augmentation are  $v_1$  and  $v_2$ .

After data augmentation,  $v_1$  and  $v_2$  extract representation vectors  $h_1$  and  $h_2$  in the encoder. The encoder can arbitrarily decide depending on the downstream tasks. In this experiment, ResNet (He et al., 2016) is chosen as the encoder.

Following feature extraction, a small neural network projection head maps representations to the space (Chen et al., 2020b) in which contrastive loss is applied. We use a 2-layer MLP with ReLU and BN layers to project the feature vectors  $h_i$  and  $h_j$  to the 256-dimensional vectors  $z_1$  and  $z_2$ . This projection head is not involved in downstream task training.

This paper proposes two methods, the classification distance and the classification distance, which combine the a priori distances of various types of image pairs with the actual distances of the corresponding vectors  $z_1$  and  $z_2$  to compute the contrastive loss. The calculation process for grading distance and classification distance will be detailed in Section 3.3.

Fig. 1 illustrates the contrast between previous contrastive learning strategies and PDE. In conventional contrastive learning, the contrastive loss is computed using different views of the same image as positive samples and other images as negative samples to determine the contrastive loss. Therefore, a large batch size or momentum encoder is required to provide more negative samples to facilitate the model's convergence. PDE computes contrast loss based on various types of image pairs without requiring a large batch size or momentum encoder, and the model can be trained to converge with 64 batch size.

### 3.2. Phenological stages acquisition and comparative learning image pair construction

The construction of image pairs involves three steps: sampling plant phenological stages, image extraction, and image pairing. After these

three steps, the prior domain knowledge of phenological stages can be transferred to four types of image pairs, each containing different prior distances.

First, we extract phenological period information from plant images. For  $n$  image sequences of a specific phenotypic research target plant, 3–5 sample sequences are chosen at random. For each selected example sequence, the bud emergence is set as the reference time day0, and the start time and duration of various phenological periods can be determined by manually interpreting the image sequence. The example sequence is averaged over multiple time intervals. This average can approximately represent the phenological period of all  $n$  image sequences due to the relative constancy of the annual growth cycles of individual plants. Fig. 2(a) depicts the acquisition of phenological periods for a selection of example sequences. This step requires manual interpretation of the phenology data, but the effort required to interpret 3–5 example sequences of phenology is significantly less than that required to classify, target, and semantically annotate large batches of plant image data. Additionally, information regarding the phenological period can be obtained from the previous agricultural research literature.

Second, the desired images are extracted from the images of different phenological periods. Although the average phenological period of the example images can be approximated to the overall phenological period, the exact dates of the phenological period intersection of various time series images are not identical. And phenological period change is a process. At the end of one phenological period and the start of the next, images may contain similar semantic information. To achieve automatic and accurate extraction of images of different phenological periods and to maximize the semantic difference of images during different phenological periods, images near the junction period of phenological stages are discarded, and only images far from the time

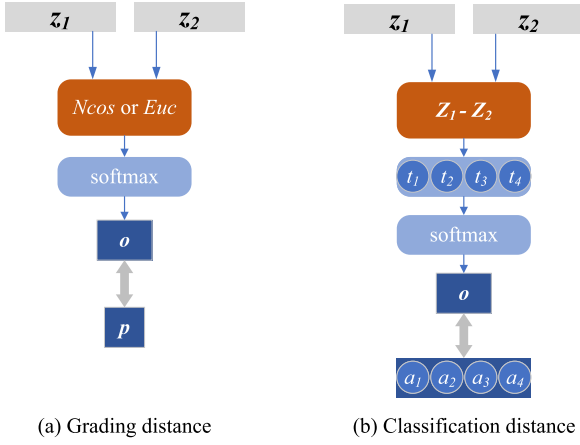


Fig. 3. Distance metric and comparison loss function.



Fig. 4. Cherry image collection device.

threshold are chosen and recorded for classification. An example of extracted images is shown in Fig. 2(b).

Third, following the acquisition of the desired images, the images must be paired, which can be viewed as the Cartesian product of the images themselves. The obtained image pairs, which are saved as labels with One-hot encoding, can be recorded as one of four types based on the sequence and phenological period of the two images: the same period of the same sequence, different periods of the same sequence, different sequences of the same period, and different sequences of different periods. So far, domain knowledge has been incorporated into deep learning data.

### 3.3. Contrastive loss functions embedding prior knowledge

Here, hierarchical distance and categorical distance are introduced to calculate the contrastive loss based on the prior distances of different types of image pairs. In contrast, traditional loss functions such as NCEloss (Hadsell et al., 2006), InfoNCE (He et al., 2020), NT-Xent loss (Chen et al., 2020b) etc., are only used to compute the loss between positive and negative examples and cannot handle distances with different categories and hierarchies.

#### 3.3.1. Grading distance loss

The concept of hierarchical distance in this paper is derived from assigning a hierarchy to the prior distances of the four types of image pairs. Apparently, image pairs from the same sequence and the same period have the closest distance, while those from different sequences and different periods have the farthest distance. The distances between pairs from the same sequence but different periods and pairs from different sequences but the same period lie in between. Hierarchical distance is based on this known knowledge and assigns distance coefficients to each category of image pairs accordingly. Fig. 3(a) shows the grading distance.

For the image pairs of class  $l$ , a distance coefficient  $k_l$  is first defined to characterize the relative distance between the pairs of images of that class. For any pair of images  $x$  and  $y$ , the distance  $p_{xy}$  can be obtained:

$$p_{xy} = \frac{k_1}{\sum_{i=1}^4 k_i} a_1 + \frac{k_2}{\sum_{i=1}^4 k_i} a_2 + \frac{k_3}{\sum_{i=1}^4 k_i} a_3 + \frac{k_4}{\sum_{i=1}^4 k_i} a_4 \quad (1)$$

$[a_1, a_2, a_3, a_4]$  represents the original labels of the image pairs. After this processing, the prior distance  $p_{xy}$  of any image pair  $x$  and  $y$  is labeled as the number between  $[0, 1]$ ,  $p_{xy}$  can be considered as the probability that the sample pair is the same semantic image.

For the vectors  $z_1$  and  $z_2$  of image pairs  $x$  and  $y$ , the loss is calculated as:

$$L_{xy} = - \left[ p_{xy} \log \frac{1}{1 + e^{\frac{sim(z_1, z_2)}{\tau}}} + (1 - p_{xy}) \log \left( 1 - \frac{1}{1 + e^{\frac{sim(z_1, z_2)}{\tau}}} \right) \right] \quad (2)$$

Here,  $Sim(z_1, z_2)$  represents the similarity between  $z_1$  and  $z_2$ . The function can be seen as calculating the similarity between  $z_1$  and  $z_2$ , followed by the Sigmoid function and cross-entropy with the probability represented by  $p_{xy}$ . The temperature coefficient, denoted as  $\tau$ , can be used to adjust the distribution of similarities. Similarity can be measured using negative cosine similarity or Euclidean distance. When using negative cosine similarity:

$$Sim(z_1, z_2) = NCos(z_1, z_2) = \frac{1}{2} \left( 1 - \frac{z_1^T z_2}{\|z_1\| \|z_2\|} \right) \quad (3)$$

Euclidean distance is the  $L_2$  norm:

$$Sim(z_1, z_2) = Euc(z_1, z_2) = \|z_1 - z_2\|_2 \quad (4)$$

For each mini-batch with a batch size of  $n$ , the contrastive loss, denoted as loss, is given by:

$$L_B = \frac{\sum_{i=1}^n L_i}{n} \quad (5)$$

#### 3.3.2. Classification distance loss

In hierarchical distance, the contrastive loss is calculated based on the similarity of image pairs and the assignment of prior distances for each category of image pairs. In contrast, categorical distance does not require explicit assignment of different distances for various image pairs. Instead, it implicitly maps the distance information of the contrastive model to a fully connected layer and directly calculates the loss with the classification of different image pairs. Fig. 3(b) illustrates the categorical distance.

When calculating the Euclidean distance between  $z_1$  and  $z_2$ , the equation can be further written as:

$$Euc(z_1, z_2) = \|z_1 - z_2\|_2 = \sqrt{(z_{11} - z_{21})^2 + (z_{12} - z_{22})^2 + \dots + (z_{1n} - z_{2n})^2} \\ = \sqrt{\sum_{i=1}^n (z_{1i} - z_{2i})^2} \quad (6)$$

When calculating the contrastive loss with categorical distance, the subtraction of  $z_1$  and  $z_2$  is performed element-wise, as follows:

$$e = z_1 - z_2 \quad (7)$$

where  $e$  represents a vector of the same dimension as  $z_1$  and  $z_2$ , and it can be found that the process of computing  $e$  is the central step in computing the Euclidean distance between  $z_1$  and  $z_2$ , which makes  $e$  contain a priori distance information of the images pairs. Projecting  $e$  linearly to a fully connected layer  $t$  with 4 nodes, processed by SoftMax, yields the output  $o$ :

$$t = eW \quad (8)$$

$$o = \text{Softmax}(t) \quad (9)$$

Cross-entropy (Rubinstein, 1999) is used to calculate the error between the category information in  $o$  and the prior distance information of the label categories for image pairs. That is:

$$L = - \sum_{i=1}^4 a_i \log_2 o_i \quad (10)$$

#### 4. Cherry phenotype temporal modeling based on PDE pre-training

Section 3 has provided a detailed description of the theoretical foundations and implementation details of the PDE method. To provide a comprehensive example of plant temporal image phenotypic analysis, this section establishes a cherry growth model based on the PDE method. Specifically, it includes cherry image acquisition, PDE pre-training, cherry temporal image anomaly detection, cherry organ semantic segmentation, and cherry growth numerical model establishment. The details are described as follows.

##### 4.1. Experimental image acquisition

In the process of plant temporal phenotype modeling, the first step is to collect plant time-series images. Without loss of generality, this paper uses the Hikvision iDS-2DC4223IW-/GLT(S5) model and iDS-2DC2204IW(S6) model cameras to build a remote image acquisition device, as shown in Fig. 4. Time-series images of the Tieton were collected from 16 different greenhouse sheds in Dalian, China, as the experimental data for this study. The cameras can be set to multiple preset points, and the pan-tilt unit can precisely capture cherry images at preset angles by rotating and zooming. Images are automatically captured every 3 h starting from 0:00 each day. A total of over 40,000 time-series images were collected during two complete growth cycles of cherries (January 2021 to July 2022).

##### 4.2. Cherry image PDE pre-training

Once the time-series images of cherries are obtained, the pre-training process can be initiated as described in the third section of this paper. The details are not reiterated here.

##### 4.3. Anomaly detection

In practical scenarios, images collected are susceptible to anomalies due to agricultural activities, as illustrated in Fig. 5. Anomalies such as insufficient lighting, occlusions, water droplets, overexposure, defocus, and blurring can result in the loss of effective information in the images. When the loss of information exceeds a certain threshold, the features extracted by the encoder become inadequate for downstream tasks such as semantic segmentation. Additionally, images with significant deviations in shooting angles compared to other images in the sequence are also considered anomalous. While such images may not suffer from the loss of effective information themselves, they can still introduce noise during later modeling stages. To address anomalies in the cherry

image sequence, this paper utilizes the backbone network pre-trained by PDE as a feature extractor, combined with a Variational Autoencoder (VAE) (An and Cho, 2015), to accomplish unsupervised anomaly detection for the time series images.

Specifically, after cherry images are input into the encoder pre-trained by PDE, the output is a 256-dimensional feature vector, which can be regarded as a point in a 256-dimensional space. Consequently, the features of the cherry image sequence can form a curve in the 256-dimensional space. It is evident that plant growth occurs slowly, and the changes in cherry temporal images over a period of time are relatively small. Corresponding to the curve in the space, it should be a smooth curve that changes within a certain range. If there are anomalous images in the sequence, the feature point corresponding to that image should deviate from the range of the curve. It can be observed that the anomaly detection for plant temporal images is essentially transformed into the detection of abnormal points on a curve in multidimensional space. Furthermore, to accomplish this task, feeding the output spatial curve into a VAE can effectively perform anomaly point detection. Specifically, the feature vector extracted by the PDE encoder is input into the Encoder of the VAE. The Encoder of VAE utilizes fully connected layers to generate a lower-dimensional vector, describing the distribution of data in the latent space to obtain latent variables  $g$ . It ultimately outputs the mean vector  $\mu$  and variance vector  $\sigma$  in the latent space. To ensure that  $\sigma$  is positive, the original values output by the network are subjected to the exponentiation transformation. The dimensions of these two vectors are the same as the dimensions of the latent space, as shown in the following formula:

$$\mu = \text{Encoder}(a) \quad (11)$$

$$\log(\mu^2) = \text{Encoder}(a) \quad (12)$$

$$g \sim N(\mu, \sigma^2) \quad (13)$$

Where  $\text{Encoder}$  is the computational function of the network, and  $N$  relates to a normal distribution with a mean of 0 and a variance of 1.

The inverse transformation of the latent space was achieved by inputting the sampled latent variable vector into the decoder of VAE. This process reconstructs the high-dimensional representation  $m'$  of the original input data. The decoder is also composed of fully connected layers, gradually increasing the number of neurons to decode the distribution information into a high-dimensional vector. Ultimately, the reconstruction of the data is completed through the output layer, resulting in the restoration of the high-dimensional vector.

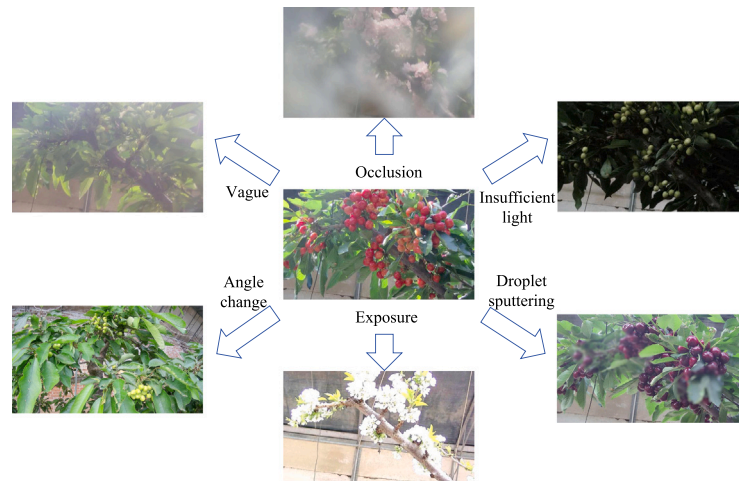
The model compares the differences between the reconstructed data generated by the decoder and the original input data. Typically, anomalous data points exhibit larger reconstruction errors, markedly different from the normal data points learned by the model. Exactly, the model achieves learning effective representations of the data by minimizing the loss function through the autoencoder. The loss function includes the reconstruction loss, typically measured using Mean Squared Error (MSE) to quantify the difference between the reconstructed image and the original image. Additionally, the Kullback–Leibler (KL) divergence is used to measure the difference between the distribution of latent variables and the standard normal distribution. The specific formula for this is as follows:

$$L_{recon} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^D (m_{ij} - m'_{ij})^2 \quad (14)$$

$$L_{kl} = \frac{1}{2} \sum_{j=1}^J (1 + \log(\frac{\sigma_j^2}{e}) - \mu_j^2 - \sigma_j^2) \quad (15)$$

$$L_{total} = L_{recon} + L_{kl} \quad (16)$$

Where  $N$  is the number of samples,  $D$  is the dimensionality of the input data,  $m_{ij}$  represents the  $j$ th dimension of the  $i$ th input sample,



**Fig. 5.** Anomaly scenario illustration. Occlusion: foreign objects blocking the lens; Insufficient light: the insulation blanket was not removed in a timely manner, resulting in the inability of the light inside the greenhouse to shine; Droplet sputtering: watering or rain dripping from the greenhouse onto the camera; Exposure: light too bright; Angle change: due to network latency, the program settings for the device's timing rotation are not in place; Vague: focusing failed.

$m'_{ij}$  is the corresponding value in the reconstructed output data,  $\mu_j$  and  $\sigma_j$  are the mean and variance output by the encoder,  $J$  is the dimensionality of the latent space,  $\epsilon$  is a small constant for numerical stability.

Through this stage of the network, abnormal images can be labeled as 1, while normal images are labeled as 0, completing the anomaly detection for cherry growth images.

#### 4.4. Semantic segmentation

This study takes the semantic segmentation of cherry phenotypic images as an example, employing the U-Net semantic segmentation network (Ronneberger et al., 2015). Pixel-level segmentation is performed on cherry images to finely segment branches, flowers, leaves, fruits, and background regions, aiming to obtain detailed phenotypic information for each image in the cherry time series.

U-Net is a classic encoder–decoder architecture that can utilize various backbones as the encoder. This flexibility makes it easy to transfer pre-trained weights from contrastive learning to the U-Net network. The encoder network can initially obtain five preliminary effective feature layers (feature maps). In the decoder network, utilizing transposed convolutions on the five feature layers, along with feature fusion, allows the acquisition of an effective feature layer that consolidates all the features. In plant image segmentation, where semantics are relatively simple and structures are fixed, the U-shaped structure's feature concatenation can effectively combine shallow features with deep features. Classifying each feature point in the final obtained feature layer allows for the acquisition of semantic segmentation results.

Fig. 6 illustrates the schematic of transferring the pre-trained encoder from PDE to U-Net. During downstream training, the encoder is initially frozen, and only the decoder is trained. This strategy is employed to prevent the disruption of encoder weights. Subsequently, the entire network is unfrozen, and fine-tuning is performed with a smaller learning rate. The segmentation performance serves as an effective metric to evaluate the effectiveness of contrastive learning pre-training.

#### 4.5. Temporal modeling

After establishing the anomaly detection and semantic segmentation models during the information extraction stage, the temporal modeling phase records the changes in plant growth over time from a temporal

perspective. This paper establishes a model to track the changes in the area proportions of different plant organs over time, including stems, leaves, flowers, and fruits. Furthermore, the model fits the richards growth curve (Richards, 1959) to describe the growth patterns. The paper also records the color changes during plant growth by documenting the average Hue (H) values of all pixels in different organ image regions of the plant. This process is used to model the color of plant organs.

##### 4.5.1. Temporal modeling of organ area

**Extracting growth curves.** Performing time series analysis on images processed through semantic segmentation, tracking the changes in the area proportions of different plant organs over time, including stems, leaves, flowers, and fruits. This process generates organ proportion area curves that vary with time.

Firstly, detailed image processing is conducted on the image data at each time point, involving image segmentation and pixel-level analysis. This study precisely calculates the pixel proportion area for each plant organ (such as leaves, fruits, etc.) and fits these discrete proportion values into curves. This is achieved by connecting adjacent discrete points with line segments to approximate the overall trend of the data. Thus, a data sequence that varies with time is obtained, encompassing the relative area proportions of different plant organs at different growth stages.

Next, the obtained organ growth curve proportion plots undergo Kalman filtering to enhance the accuracy and reliability of the data. State estimation of time series data involves updating by merging prior information and observational information, while considering noise and uncertainty. This enables precise tracking of dynamic changes in the proportion of different organs over time, without being affected by measurement errors.

In plant growth, the growth of organs is a gradual process without repeated fluctuations in organ size. Applying monotonicity processing to the curves ensures that the results better align with the actual process of plant growth. To address occasional unreasonable fluctuations at individual time points, a monotonicity processing method based on comparing adjacent data points is employed. That is, if the value at the subsequent time point is greater than that at the preceding time point, the value at the subsequent time point is adopted; conversely, if the value at the subsequent time point is less than that at the preceding time point, the value at the preceding time point is retained. Through monotonicity processing, it ensures that the curve exhibits a monotonic increasing or decreasing trend over time.

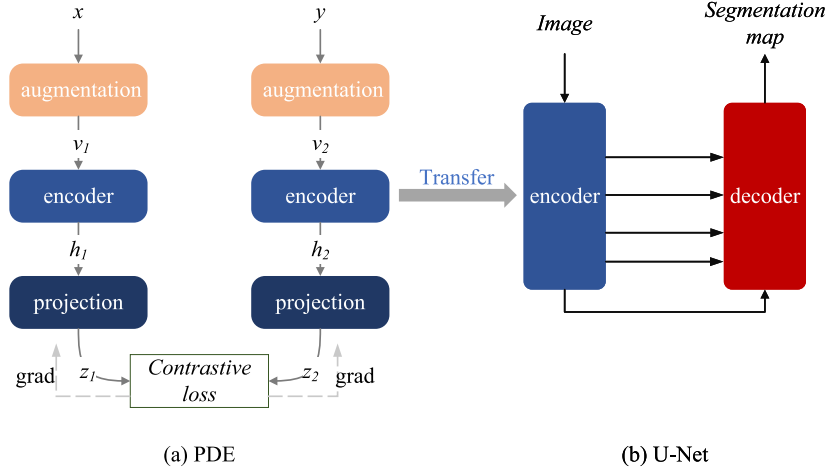


Fig. 6. Pre-training and transfer.

**Fitting growth value models.** Based on data parameterized growth curves, the model transitions from nonnumeric to numeric, aiding in quantifying the plant's growth process. Richards model (Richards, 1959) can effectively capture the saturation and gradual deceleration characteristics of biological growth, reflecting the plant growth process.

The fitting formula for the richards model (Richards, 1959) is as follows:

$$V(t) = \frac{K}{(1 + e^{-b(t-T_0)})^{\frac{1}{v}}} \quad (17)$$

Where  $V(t)$  is the value of the growth variable at time  $t$ ,  $K$  is the maximum growth value representing the limit that the growth curve eventually tends toward,  $b$  is the growth rate parameter controlling the speed of growth,  $t$  is the time point of the observed data point,  $T_0$  is the growth start time or time offset representing the starting point of the curve, and  $v$  is the shape parameter controlling the shape of the curve.

Curve fitting is based on minimizing the error function and non-linear parameter optimization to find the optimal parameter estimates. This is done to fit the richards curve model to achieve the best match with the curves obtained after the aforementioned processing. The fitting process is essentially an iterative optimization problem. The task of the fitter is to minimize the error by adjusting the parameters of the richards function model, achieved through the use of optimization algorithms. This study provides initial parameter guesses for the model. The algorithm automatically estimates initial parameter values based on the characteristics of the observed data and the form of the model, offering a reasonable starting point. Subsequently, the fitting process further refines these parameters, often denoted as  $(K_0, b_0, t_0, v_0)$ . The Levenberg–Marquardt algorithm (Ranganathan, 2004) is iteratively used to adjust parameters to minimize the sum of squared residuals, gradually reducing the error until the best match is achieved. This process can be iterated multiple times until a predetermined convergence criterion is met or the maximum number of iterations is reached. This study uses the fitting coefficient  $R^2$  (Karadavut et al., 2010) to measure the degree of fit of the fitting model to the observed data, with the formula as follows:

$$R^2 = 1 - \frac{SSR}{SST} \quad (18)$$

$$SSR = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (19)$$

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (20)$$

Where  $SSR$  is the sum of squared residuals, representing the sum of the squared differences between the predicted values of the model

and the actual observed values,  $SST$  is the total sum of squares, representing the variance of the observed data,  $n$  is the number of observed data points,  $y_i$  represents the actual observed value,  $\hat{y}_i$  represents the predicted value of the model for the  $i$ th observation, and  $\bar{y}$  represents the mean of the observed data.

In particular, due to the presence of two swelling periods in fruits, this study divides the entire growth curve into two segments at each time point. The richards curve fitting model is then applied separately to each segment. By enumerating all possible segmentation points from the first time point to the last time point, calculating the curve fitting result  $R^2$  for each point on both sides of the segmentation, the optimal segmentation point is sought.

#### 4.5.2. Temporal modeling of organ color

Recording the color information of leaves, flowers, and fruits in the images that have been semantically segmented over time helps to explore the growth patterns and developmental trends of different organs' colors more deeply.

In computer vision, HSV (Hue-Saturation-Value) is a commonly used color space, where the hue (H value) represents the basic color tone of the image. Its value typically ranges from 0 to 360 degrees, with red at 0 degrees, green at 120 degrees, and blue at 240 degrees. This representation of colors has cyclic properties, making color changes more intuitive and distinguishable.

In this study, the images obtained from various organ regions are converted to the HSV color space. By analyzing the H value, color information can be effectively separated without being affected by brightness and saturation, enabling accurate capture of the color characteristics of different plant organs. The H value of each pixel in each organ of each sequence is calculated, and then the average H value of each organ is computed and accumulated over time. This approach yields a growth curve model based on color values, reflecting the growth status and color change trends of plant organs.

## 5. Experiment and analysis

This chapter, based on cherry temporal images, demonstrates the effectiveness of PDE pre-training and presents a comprehensive growth modeling process.

During contrastive learning training of the encoder, meaningful data representations were learned through proxy tasks. In contrastive learning, the performance of the pre-trained model is typically evaluated by fine-tuning and testing on downstream tasks. In the experiment, the contrastive learning pre-training of plant temporal images with embedded prior distance measures was conducted in three parts as outlined in this paper. To assess the performance of PDE pre-training, this



Fig. 7. Cherry image sequences.

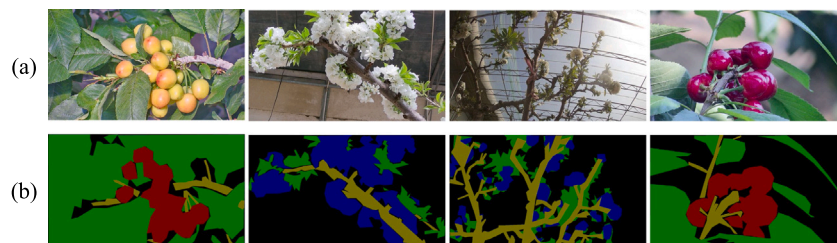


Fig. 8. Cherry images along with semantic segmentation labels for each organ. (a) Cherry color images. (b) Cherry organ segmentation labels.

paper conducted ablation experiments and comparative experiments on downstream tasks, specifically in cherry phenotypic image semantic segmentation. It is important to note that in general applications, the semantic segmentation of plant organs should occur after pre-training and anomaly detection steps.

After validating the pre-training effects and obtaining the pre-trained encoder, the cherry temporal model was established following the four parts described in this paper.

### 5.1. Experimental data

The experimental images were obtained from the experimental setup described in Section 4.1. In the study of contrastive learning methods for plant temporal images with embedded prior distance measures, 21 cherry growth image sequences (numbered 1 to 21) collected in 2021 were selected for pre-training, totaling 7373 images. An example image sequence is shown in Fig. 7. To ensure the robustness of the algorithm in practical use, real collected image data was utilized in this experiment.

In the sequence of images, 475 images were randomly selected for semantic labeling, with 335 in the training set, 70 in the test set, and 80 in the validation set. Stem, leaf, flower, and fruit were segmented in the cherry images to serve as training and testing data for semantic segmentation, as shown in Fig. 8.

In the temporal modeling stage, 20 sequences of cherry images collected in 2022 were selected as the subjects for modeling. In the temporal modeling stage, 20 sequences of cherry images collected in 2022 were selected as the subjects for modeling.

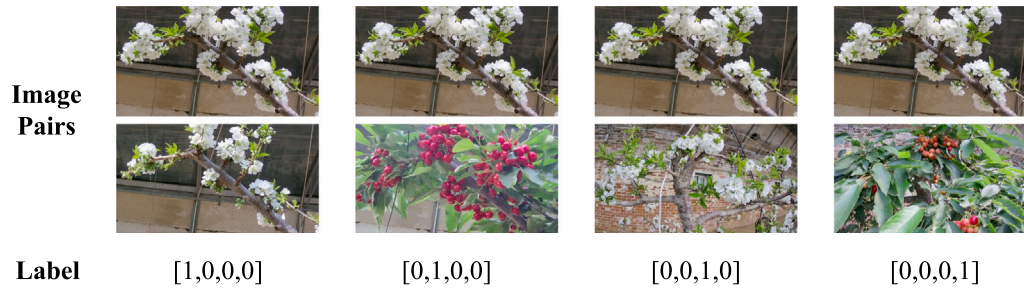
### 5.2. Training parameters and environment

All experiments adhere to the following settings. PDE employs ResNet-50 (He et al., 2016) as the encoder, with an input image size of  $512 \times 288$ , a batch size of 64, and training for 10 epochs. RMSprop (Tieleman et al., 2012) is utilized as the optimizer with a learning rate of 0.001 and a discounting factor of 0.9.

Evaluate its quality through downstream semantic segmentation tasks based on U-Net (Ronneberger et al., 2015) for plant phenotypic image. The U-Net network is an encoder–decoder structure, where the encoder is transferred from a contrastive learning pre-trained ResNet50 network, and the decoder uses Glorot Uniform initialization (Glorot and Bengio, 2010). The dataset comprises 475 images, with 335 in the training set, 70 in the test set, and 80 in the validation set. The model's input image size is  $1024 \times 512$ . During training, the encoder is initially frozen, and only the decoder is trained to prevent the disruption of encoder weights. Subsequently, the entire network is unfrozen, and the network is fine-tuned with a smaller learning rate. Training utilizes Adam (Kingma and Ba, 2014) as the optimizer. Focal loss (Lin et al., 2017) and Dice loss (Milletari et al., 2016) are employed for training. Focal loss helps mitigate sample imbalance, while Dice loss is more suitable for use in semantic segmentation tasks. During frozen training, the initial learning rate is set to 0.0001, the batch size is 18, and the model is trained for 60 epochs. During fine-tuning, the initial learning rate is set to 0.00001, the batch size is 8, and the model is trained for 60 epochs. The learning rate is linearly decayed with a decay rate of 0.96. For segmentation tasks, IoU (known as Jaccard index Jaccard, 1912)

**Table 1**  
Cherry phenology stage division and image extraction(day).

No.	Cherry phenology				Image extraction
	3	5	11	avg	all
Dormancy	-14~0	-14~0	-11~0	-13~0	-12~0
Building and flowering	1~32	1~32	1~34	1~33	10~22
Fruiting	33~73	33~67	35~75	34~72	45~57
Ripening	74~117	68~102	76~105	73~108	84~96



**Fig. 9.** Cherry images pairs.

and PA (Pixel Accuracy) for each class are computed on the validation set. Additionally, mIoU and mPA for all classes are calculated.

For the anomaly detection task, the VAE encoder hidden layer consists of three MLP layers with neuron counts [1, 4, 1], and the VAE decoder hidden layer consists of three MLP layers with neuron counts [4, 4, 4] (An and Cho, 2015).

All experiments were conducted on an NVIDIA A6000, using the TensorFlow and Keras deep learning frameworks. The tods framework was employed for anomaly detection. All results are averages from 5 experiments.

### 5.3. PDE pre-training performance verification and semantic segmentation

#### 5.3.1. Image pair construction

To train the PDE network encoder, the contrastive learning image pairs are constructed according to the method outlined in Section 3.2.

First, obtain the phenology of cherries. Sequences No. 3, No. 5, and No. 11 are drawn as example sequences. The Bud emergence was set as the reference time day0, and the phenological period was obtained by manually interpreting the image sequence, including the dormant period, flowering and budding, fruit setting, and ripening. The records are shown in Table 1 Cherry phenology. Although the time nodes of the beginning and ending of the cherry growth period are distinct, they generally remained consistent.

Second, the cherry images that are desired are extracted. Twelve images were chosen for each growth period of each sequence. The images of the period close to the phenological period's boundary were discarded, and the images of the period far from the time threshold were chosen. Details of the selection time and its corresponding species are shown in Table 1 Image extraction.

Third, after acquiring the desired cherry images, the images are paired. As depicted in Fig. 9, the records saved as labels with One-hot encoding were classified into four groups: the same period of the same sequence, different periods of the same sequence, different sequences of the same period, and different sequences of different periods.

When constructing image pairs, 3310 experimental images are extracted from 7373 images of the 21 cherry growing image sequence. By constructing images pairs in pairs, 56245 images pairs of the same sequence and same period, 219597 images pairs of the same sequence and different period, 1051553 images pairs of different sequences and different periods, and 4149000 images pairs of different sequences and different period could be obtained. To balance the training samples, data resampling was performed for different types of image pairs (Hu et al., 2022), with the sampling set to 100,000 or 150,000 for each class of image pairs, depending on the experiment.

**Table 2**

Grading distance and classification distance calculation.

Methods and parameters	mIoU	mPA
Grading distance 1.0,0.4,0.8,0.0	0.525	0.6666
Grading distance 0.8,0.4,0.6,0.2	0.5234	0.665
Grading distance 1.0,1.0,0.0,0.0	0.5074	0.6451
Classification distance	0.5675	0.7182

#### 5.3.2. Grading distance and classification distance calculation

This study trained on image pairs to investigate the impact of two contrastive loss metrics that included a priori distances, graded distance, and classification distance. According to intuitive experience, the distances between image pairs increase from near to far as follows: same sequence, same period, same sequence, different period, and different sequence, different period. On the basis of this intuitive experience, two sets of distance coefficients (in the order of same sequence same period, same sequence different period, different sequence same period, different sequence different period, the same below) 1.0, 0.4, 0.8, 0.0 and 0.8, 0.4, 0.6, 0.2 were chosen as comparisons, and the cases in which the distance coefficients were 1.0, 1.0, 0.0, 0.0 were also tested, which degenerated to only positive and negative examples are compared for contrast learning. There is no need to define coefficients manually for classification distances.

After migrating to the segmentation task, Table 2 compares the impact of grading distance versus classification distance. Only the positive and negative examples, with distance coefficients of 1.0, 1.0, 0.0, and 0.0, are shown to be the least effective for training, indicating that grading distance and classification distance are effective contrastive loss measures. The classified distances are superior to the distances graded with two distinct coefficients. It is demonstrated that implicitly obtaining distance information directly from the projection and differentiating distance classes is preferable to manually determining distances and assigning values. Notably, the grading distance is an intuitive and fine-grained measure of distance. In this experiment, it may be less effective than the classification distance because the optimal distance coefficient has not been identified. In addition to particle swarm algorithms, Bayesian optimization, and other heuristics, this coefficient can also be obtained using these methods. However, there is no doubt that the classification distance is simple to calculate and effective.

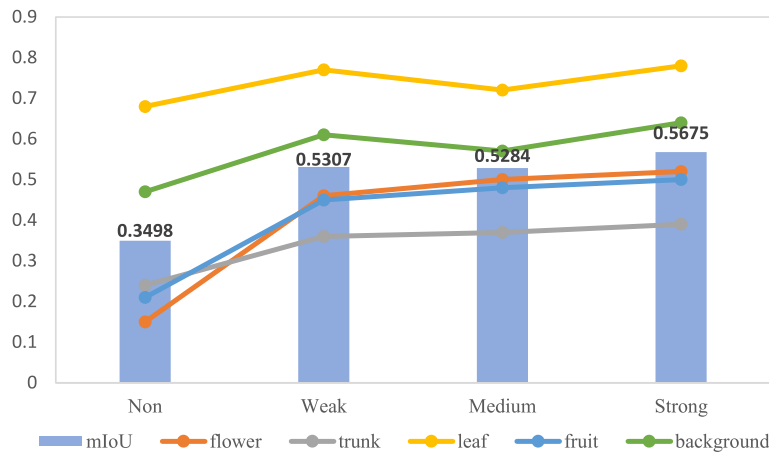


Fig. 10. Comparison of different intensity data enhancement.

Table 3

Data enhancement results.

	mIoU	Flower	Trunk	Leaf	Fruit	Background
Non	0.3498	0.15	0.24	0.68	0.21	0.47
Weak	0.5307	0.46	0.36	0.77	0.45	0.61
Medium	0.5284	0.5	0.37	0.72	0.48	0.57
Strong	0.5675	0.52	0.39	0.78	0.5	0.64

### 5.3.3. Data augmentation

After inputting image pairs into the model, data augmentation is performed to draw the model's attention to higher-level semantic information and better utilize similar images. In this paper, the employed data augmentation methods primarily include random cropping and flipping, random brightness, random contrast, random saturation, random hue, and random black and white. The experiments were designed to compare the impact of four different levels of data augmentation on the results. The data augmentation is implemented based on the image library in TensorFlow, and different coefficients can be adjusted to generate different data augmentation effects. Specifically, the coefficients for no data augmentation are 0, 0, 0, 0 (random brightness, random contrast, random saturation, random hue, respectively). For weak data augmentation, the coefficients are 0.3, 0.2, 0.2, 0.1. For moderate data augmentation, the coefficients are 0.5, 0.4, 0.4, 0.3. For strong data augmentation, the coefficients are 0.7, 0.6, 0.6, 0.5. In addition, the coefficients for random cropping and flipping, as well as random black and white, are consistent across the experiments. Classification distance is employed as the contrastive loss in this experiment. The model is trained using resampled data with 100,000 image pairs for each class. To better illustrate the effects of data augmentation, the second convolutional layer of the second residual block in the encoder ResNet50 model trained with different data augmentation levels was visualized (Fchollet, 2020). Generate input images to maximize the activation of specific filters in the target layer. Such images represent visualizations of the patterns to which the filters respond.

The effects of training with different levels of data augmentation on downstream segmentation tasks are presented in Table 3. Comparative results and IoU for each class are shown in Fig. 10. The convolutional visualization figures are depicted in Fig. 11.

In the convolutional visualization figure with no data augmentation, as shown in Fig. 11(a), the convolutional kernel texture appears smooth, and large areas of green and red are evident. This suggests that the model has learned a significant but singular color information. The convolutional kernels seem to have overlooked other important semantic information such as edges and textures. This phenomenon can be attributed to the fact that, during the plant growth process, leaves and fruits often constitute the main visual components. Consequently,

there are large areas of red and green in the images, and the extensive color coverage can impact the network's training. Reflected in the segmentation results, the overall mIoU is relatively low, and the IoU for leaves is significantly higher than for other parts.

In the convolutional visualization figures with added data augmentation (Fig. 11(b)–(d)), it can be observed that data augmentation helps alleviate the model's tendency to over-converge to color information. The convolutional kernels can now learn more advanced semantic information, such as textures. Reflected in the segmentation results, the performance with data augmentation is significantly better than without data augmentation. The strongest data augmentation yields the best results, and the IoU for each part is the highest among the different augmentation levels. The performance with weak data augmentation is similar to that with moderate data augmentation. The mIoU with weak data augmentation is slightly higher than with moderate data augmentation, but for trunk, flower, and fruit, moderate data augmentation achieves better IoU than weak data augmentation.

In summary, applying data augmentation in PDE is essential.

### 5.3.4. Comparison with other pretraining methods

In this paper, we evaluate their quality by applying downstream tasks to models with different initializations: the Glorot Uniform initialization (Glorot and Bengio, 2010), the self-supervised pre-training SimCLR (Chen et al., 2020b), MoCo (He et al., 2020), SimSiam (Chen and He, 2021), supervised training ImageNet (Deng et al., 2009) weights, the weights obtained from the supervised classification of plant phenological periods, and the grading distance or classification distance for PDE. Table 4 illustrates the effects of various initialization models on downstream segmentation tasks.

Glorot Uniform initialization (Glorot and Bengio, 2010) is TensorFlow's default method of initialization and represents the case without pre-training. In the experiments, the same freezing and fine-tuning techniques as the other methods were evaluated, with scratch training also being evaluated.

SimCLR (Chen et al., 2020b), MoCo (He et al., 2020), SimSiam (Chen and He, 2021) all use 7373 images of 21 sequences that were not extracted using the method described in Section 3.2. The comparative learning loss variation in pre-training is shown in Fig. 12. SimCLR converges poorly during training, and its transfer results on downstream tasks are only comparable to those of Glorot Uniform initialization (Glorot and Bengio, 2010) in the absence of pre-trained weight loading. Loss of MoCo decreases for a time, but subsequently increases and fails to converge. SimCLR and MoCo are both contrastive learning models that utilize both positive and negative examples. In plant time series images, there are a large number of similar images, making it difficult for the model to determine whether the same image is a positive sample enhanced by different data or a negative sample from different but

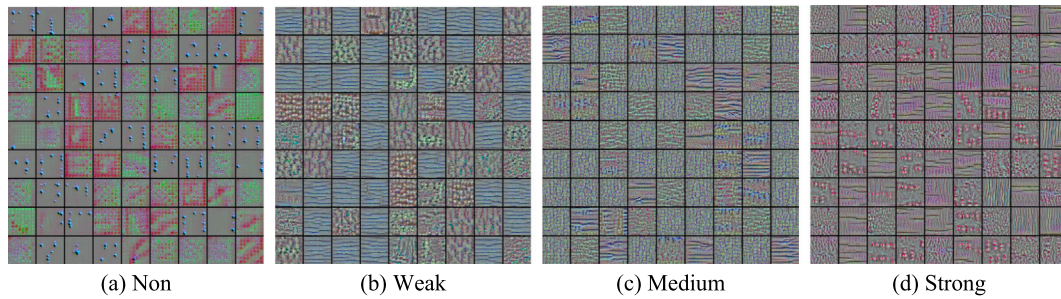


Fig. 11. Convolutional visualization with different data enhancements.

Table 4  
Result of method comparison.

	Pre-training data set	Frozen		Fine-tuning	
		mIoU	mPA	mIoU	mPA
Glorot Uniform		0.2309	0.3183	0.2652	0.3758
Glorot Uniform (from scratch)				0.2719	0.3666
SimCLR	7373 cherry images	0.0733	0.2	0.2678	0.382
MoCo	7373 cherry images		Pre-training non convergence		
SimSiam	7373 cherry images	0.2252	0.3146	0.2609	0.347
ImageNet	More than 10000000 images	0.2969	0.4158	0.5412	0.716
Supervised phenological classification	3100 cherry images	0.343	0.4502	0.3883	0.498
PDE (classification distance)	600000 image pairs from 3100 cherry images	0.5502	0.7123	0.5751	0.7355

similar images. SimCLR can converge on a poor result, while MoCo cannot converge. One of the explanations is that the SimCLR in this experiment is limited by the hardware and does not use the large batch size (4096) described in the original paper, which decreases the likelihood of a mini-batch containing images that are extremely similar. In contrast, the momentum encoder in MoCo enables the model to collect data from all other images in a mini-batch with a smaller batch size. The training increases the amount of similar image data contained in the momentum encoder. The greater the quantity, the loss decreases and then increases. SimSiam quickly converges during training, but based on the results of transferring downstream tasks in Table 4, as a comparison learning model with only positive examples, it collapses during training. The previous methods do not account for the timing and characteristics of plants and are, therefore, unsuitable for the contrastive learning of images of plant timing.

ImageNet (Deng et al., 2009) contains one million images with over 20,000 categories and is the most popular classification dataset currently trained to obtain supervised pre-trained weights that are widely migrated across multiple domains. The effect of migrating ImageNet (Deng et al., 2009) weights was poor during freezing training and improved following fine-tuning, but its overall effect was inferior to that of PDE. It demonstrates that general ImageNet image features, such as edges and corner points, are not identical to plant images. In addition, ImageNet data volume reaches millions and is not easy to train. PDE is not only a method for generating improved weight initialization, but also provides design flexibility for encoder networks.

Supervised Classification of Plant Phenological Stages. As described in Section 4.1, before generating image pairs, the extracted images are assigned a phenological stage. Traditional supervised training is performed on the ResNet50 network, and its weights are transferred. The classification accuracy reached 0.94, but its performance is less satisfactory when transferred to downstream segmentation networks. The results indicate that the features learned by the network for direct classification differ significantly from the features required for segmentation, and fine-tuning is also unable to adjust them accordingly. In this experiment, it is considered that semantic segmentation operates at the pixel level, and the encoder needs to extract more fine-grained feature information compared to a classification network.

The PDE hierarchical distance and the classification distance methods, as described earlier, were applied using resampled data with

150,000 pairs for each category. To compare various pre-training methods, the same data augmentation was applied to other contrastive learning methods, and supervised phenophase classification also had the same data augmentation added at the input end. The pre-training loss variation is shown in Fig. 12. The results of frozen training with PDE are significantly better than other methods, indicating that the encoder trained with PDE can effectively extract information from the plant. The performance is further improved after fine-tuning. Fig. 13 shows the segmentation results of this model on the cherry image sequence illustrated in Fig. 7. Overall, PDE is the most suitable pre-training model for semantic segmentation tasks on plant time-lapse images.

#### 5.4. Abnormal detection performance verification

To enhance the anomaly detection process, this study meticulously selects sequences from the complete growth cycle of cherry plants. These sequences span the germination, flowering, fruit growth, maturity, and harvesting periods. The methodology, as detailed in Section 4.3, aims to identify anomalous images that suffer from various quality issues, including insufficient lighting, occlusion, improper angle rotation, the presence of water droplets, overexposure, and issues related to defocus and blurriness. Through this meticulous process of anomaly identification, we compute key performance metrics such as Accuracy, F1-score, Precision, and Recall. These metrics serve to evaluate the algorithm's accuracy and robustness effectively. The results of these experiments are comprehensively presented in Table 5, offering insights into the effectiveness of our approach in managing image quality challenges within the dataset.

From the above table, it can be seen that the PDE pre-training method and the combination of VAE applied to cherry image anomaly detection yield the best results. The fusion of these two methods demonstrates excellent performance, capable of comprehensively and accurately identifying anomalous images with issues such as overexposure and foreign objects. This implies that in the scenario of real-time collection of large-scale image data, this study is capable of reliably detecting and annotating anomalous images, ensuring the reliability of subsequent analysis and modeling tasks, and enhancing the accuracy of research outcomes.



Fig. 12. Contrastive learning method training loss.

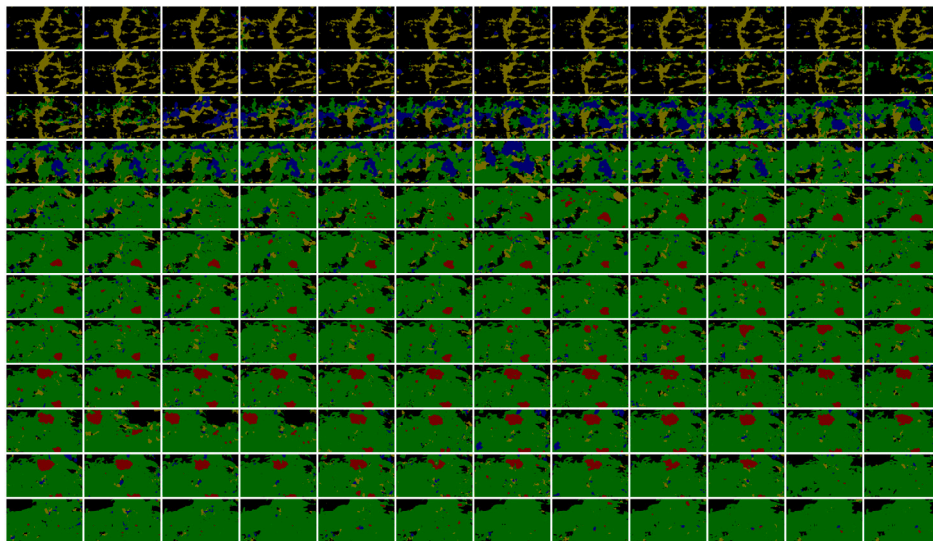


Fig. 13. Semantic segmentation of image sequences.

As demonstrated by Fig. 14, this research adeptly pinpoints images plagued by significant quality concerns. Specifically, it identifies an anomalous image within the sequence on the second day of the leaf growth curve. By effectively filtering out such anomalies, this methodology ensures the exclusion of problematic images, thereby furnishing subsequent experiments with a dataset devoid of anomalies. This capability is instrumental in detecting low-quality images, thus providing essential support for data cleansing efforts and enhancing the overall quality of analytical endeavors.

5.5. Time series modeling results

5.5.1. Temporal variation of area

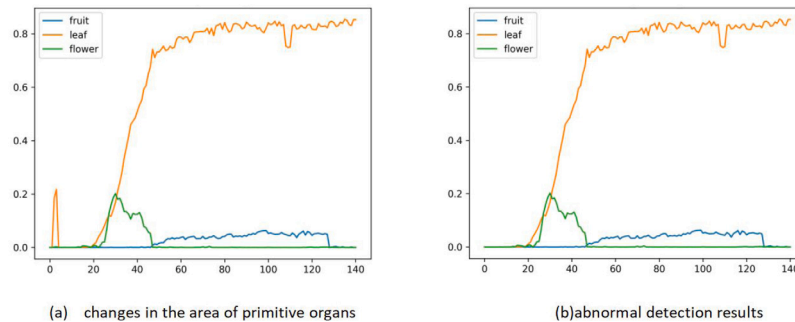
As mentioned in 4.5.1, the area of stems, leaves, flowers, and fruits is accumulated and calculated in chronological order to obtain the corresponding number of pixels for the segmented map image. The ratio curve of the pixels of each organ of the cherry plant to the entire image

pixel at the preset point 15 of the 6th camera in the E88569964 image sequence is shown in Fig. 15. The plot takes time as the horizontal axis and the proportion of organ area as the vertical axis, with values ranging from 0% to 100%. Record the relative area of different organs as they change with plant growth. The generation of a curve of the proportion area over time can provide a detailed analysis of cherry growth and organ development, providing a valuable data foundation for plant phenotype research. Due to the small changes in the stem during each growth cycle, only the leaves, flowers, and fruits are curve modeled here. Further fitting was performed to obtain the modeling results of cherry plant leaf, flower, and fruit growth. Nine sequences were selected to display the modeling results, as shown in Table 6.

By applying Kalman filtering and monotonic processing, smooth curves are obtained, reflecting the growth trends of cherry organs. The graph elucidates that, commencing on the 20th day, a simultaneous onset of growth in both leaves and flowers signifies the inception of the developmental phase. The fruit growth trajectory is delineated

**Table 5**  
Method combination results.

(a) Accuracy					
	VariationalAutoencoder	Autoencoder	IsolationForest	AutoRegODetector	DeepLog
PDE	0.9694	0.9221	0.9680	0.9156	0.9265
ResNet	0.9666	0.9185	0.9664	0.9122	0.9194
MoCo	0.9537	0.9098	0.0605	0.9068	0.9605
SimSiam	0.9689	0.0232	0.9682	0.9166	0.9264
SimCLA	0.9539	0.8946	0.9539	0.9020	0.8960
(b) F1-score					
	VariationalAutoencoder	Autoencoder	IsolationForest	AutoRegODetector	DeepLog
PDE	0.4766	0.4152	0.4568	0.3772	0.4385
ResNet	0.4233	0.3458	0.4201	0.3630	0.4045
MoCo	0.0019	0.3458	0.0000	0.3266	0.0000
SimSiam	0.4674	0.4246	0.4661	0.3891	0.4428
SimCLA	0.1753	0.2459	0.1841	0.2962	0.2529
(c) Precision					
	VariationalAutoencoder	Autoencoder	IsolationForest	AutoRegODetector	DeepLog
PDE	0.3041	0.7903	0.2770	0.7984	0.3311
ResNet	0.2846	0.6694	0.2649	0.6532	0.3046
MoCo	0.2588	0.0000	0.2345	0.0296	0.0000
SimSiam	0.2920	0.7258	0.2731	0.7500	0.3012
SimCLA	0.3146	0.7500	0.2921	0.7742	0.3425
(d) Recall					
	VariationalAutoencoder	Autoencoder	IsolationForest	AutoRegODetector	DeepLog
PDE	0.7606	0.3802	0.6984	0.3740	0.8023
ResNet	0.7613	0.3322	0.7028	0.3234	0.7912
MoCo	0.6935	0.0000	0.6356	0.0093	0.0000
SimSiam	0.7322	0.3712	0.7087	0.3691	0.7519
SimCLA	0.7872	0.3712	0.7573	0.3848	0.8420



**Fig. 14.** The results of anomaly detection.

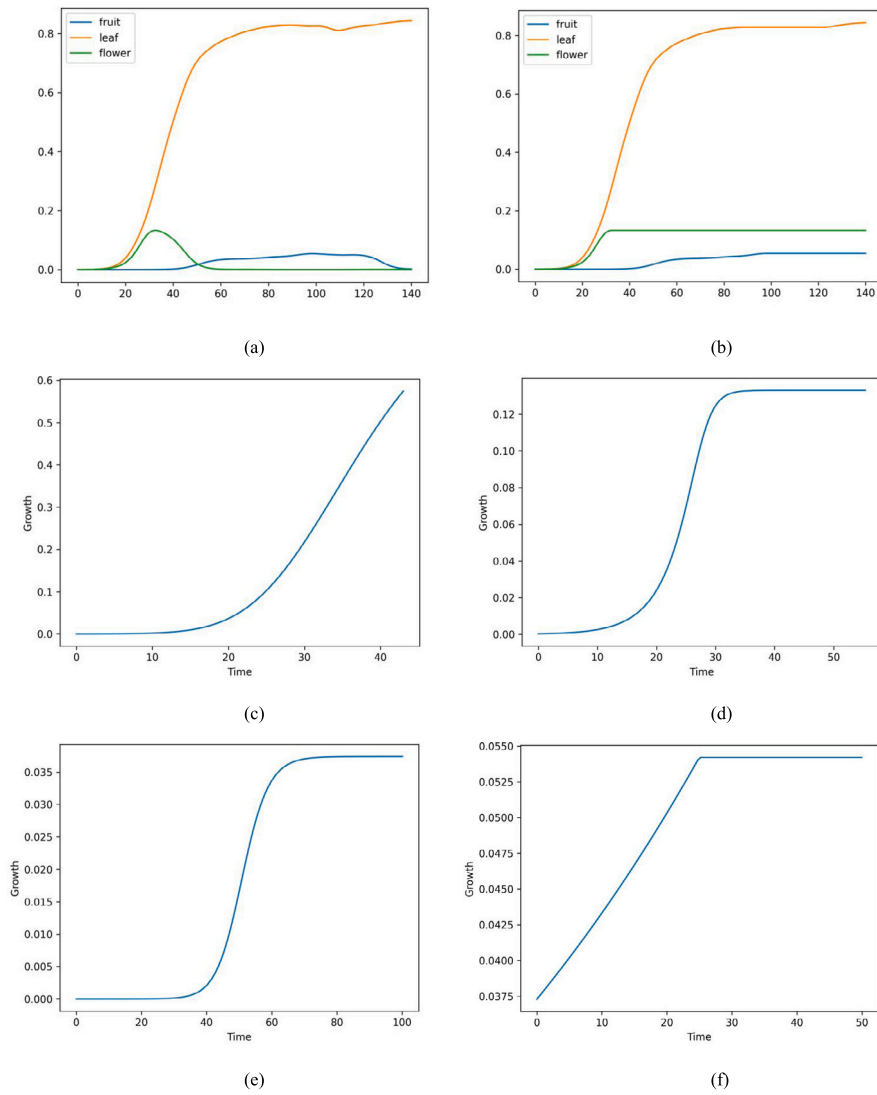
into two distinct curves, identifying the optimal segmentation point around the 100th day, which delineates the fruit's expansion phases. Notably, around the 40th day, there is a marked surge in fruit size, heralding the first expansion phase. Subsequently, by the 100th day, an observable deepening in fruit coloration coupled with a second volumetric increase underscores the presence of two pivotal expansion intervals in the cherry growth cycle. This also highlights the transformations undergone by leaves, flowers, and fruits across various stages of growth.

The coefficient of determination, denoted as  $R^2$ , possesses a value range from 0 to 1. A result nearing 1 signifies a model's superior capability to elucidate the data's variability and achieve an enhanced fit. Conversely, a result approaching 0 indicates a diminished interpretative capacity and fitting efficacy of the model regarding the data. The findings reveal a relatively steady growth rate for leaves, devoid of notable stage-wise development, with the  $v$  value approaching 1. Typically, the fruit exhibits a swifter growth rate during its initial

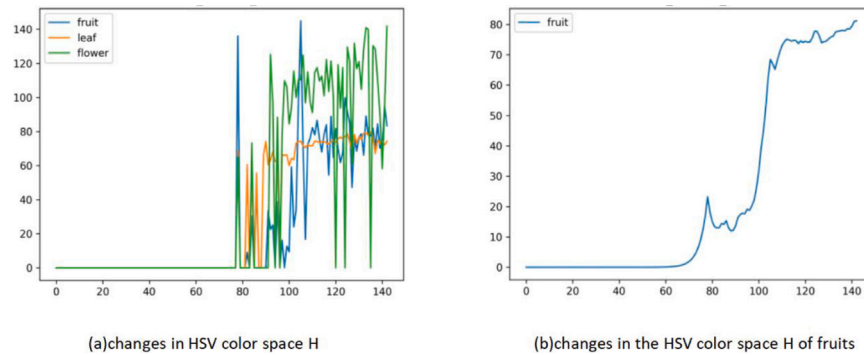
expansion phase, and the terminal growth limit in the second expansion phase notably surpasses that of other organs, reflected in a larger  $K$  value. The fitting accuracy for both leaf and fruit models across the two stages closely approaches 1, as indicated by the high  $R^2$  values. This suggests that the growth curves derived from Richards modeling are nearly congruent with the actual growth trajectories, demonstrating the model's successful adaptation to the growth trends of leaves and fruits, and its robust explanatory power concerning the data. Although the flower growth model's  $R^2$  value is lower in comparison to other cherry organ models, it still presents quite commendable fitting outcomes, suggesting that the fit for flower growth is similarly proximate to the actual growth dynamics.

#### 5.5.2. Color timing

As delineated in Section 4.5.2, this research extracts the color attributes of flowers, leaves, and fruits from the segmentation outcomes, meticulously documenting the chromatic evolution of cherry plant organs over time in a chronological series. Fig. 16 serves as an illustrative



**Fig. 15.** Area time series modeling results. (a) Kalman filtering results. (b) monotonic processing results. (c) fitting Richards curve to the area change of leaves. (d) fitting Richards curve to the area change of flowers. (e) fitting the Richards curve for the area change of fruits first segment. (f) fitting the Richards curve for the area change of fruits second segment.



**Fig. 16.** Color temporal modeling results.

example, showcasing the modeling of the chromatic timeline for cherry growth. The graph plots time along the horizontal axis against the hue values of the organs on the vertical axis, which span from 0 to 360

degrees. This visual representation effectively elucidates the dynamic color transformations of leaves, flowers, and fruits throughout their development stages.

**Table 6**  
Modeling results of cherry organ growth.

(a) Leaf fitting richards curve					
	v	b	t	k	R <sup>2</sup>
1	0.899963	0.203232	16.020463	0.769073	0.999665
2	0.903629	0.412730	7.154698	0.198799	0.997878
3	0.713203	0.196571	39.100913	0.741570	0.998802
4	0.705833	0.175624	39.299959	0.861141	0.999801
5	0.714629	0.203211	38.558335	0.703779	0.998430
6	0.795222	0.089146	17.935811	2.028488	0.999498
7	0.811377	0.122155	14.414542	1.251080	0.999644
8	0.811686	0.111870	15.619554	1.385570	0.999418
9	0.810568	0.140986	10.803979	1.202365	0.999094

(b) Flower fitting richards curve					
	v	b	t	k	R <sup>2</sup>
1	0.034920	0.193475	14.436493	3.935804	0.999793
2	0.100118	0.115389	8.978287	6.256734	0.999753
3	0.234645	0.214326	28.076804	2.277919	0.999866
4	0.216926	0.231486	26.777115	1.996094	0.999911
5	0.234542	0.214313	28.076092	2.279572	0.999867
6	0.158812	0.114156	6.923374	8.388072	0.999712
7	0.155909	0.099460	7.194434	10.656307	0.999698
8	0.163411	0.103998	7.041523	9.797716	0.999740
9	0.211402	0.042620	6.617490	34.220568	0.999551

(c) Fruit_front fitting richards curve					
	v	b	t	k	R <sup>2</sup>
1	0.001601	0.190014	8.073236	12.852492	0.998488
2	0.903629	0.412730	7.154698	0.198799	0.997878
3	0.097333	1.802048	32.120925	0.084026	0.998876
4	0.086269	0.405667	45.078504	0.500622	0.999724
5	0.097333	1.802048	32.120925	0.084026	0.998876
6	0.058449	0.531268	16.573229	0.259836	0.998396
7	0.036882	0.243325	27.426468	1.441256	0.997989
8	0.049710	0.748562	15.453162	0.208150	0.998304
9	0.023539	0.954571	11.214472	0.108436	0.999870

(d) Fruit_rear fitting richards curve					
	v	b	t	k	R <sup>2</sup>
1	0.022748	0.191174	13.727759	2.171950	0.999210
2	0.903629	0.412730	7.154698	0.198799	0.997878
3	0.235957	0.042524	20.826366	11.377787	0.998184
4	0.232781	0.034907	31.129597	62.719474	0.994080
5	0.241910	0.042572	21.388276	11.455878	0.998581
6	0.110449	0.033492	22.254583	3.563878	0.980987
7	0.100785	0.021840	49.124613	4.175098	0.986399
8	0.105253	0.050608	14.253546	2.296330	0.990396
9	0.109452	0.042383	43.592417	4.127885	0.996970

The analysis of the color time series curve derived from cherry image modeling reveals an initial H value of zero, indicative of the absence of leaves, flowers, and fruits on the cherry tree, and consequently, no observable color variations within the image. As the cherry tree matures, color changes begin to manifest across its various organs, marking the onset of growth and development.

It can be observed that due to the white color of the flowers, their color in the image is mainly influenced by the ambient light, leading to significant fluctuations in the extracted flower color. The color of the leaves tends to stabilize after experiencing a period of change over time.

To achieve a clearer representation of the color variation of fruits, apply a Kalman filter to the time series curve of fruit colors, as shown in Fig. 16(b). The significant color changes during the two fruit swelling periods of cherries are well reflected. During the initial expansion phase, the primary change observed is an increase in fruit volume, with minimal alterations in color. Conversely, the second expansion phase

is characterized by a pronounced deepening of color, reflected in a substantial rise in the H value on the curve.

## 6. Discussion

This study introduces a comparative learning approach for analyzing temporal images of plants, incorporating prior distance measurements. Experimental results demonstrate that models utilizing a classification distance loss function exhibit superior performance. The incorporation of robust data augmentation strategies within the model effectively mitigates the risk of the model solely focusing on color representation, thus preserving crucial semantic details like edges and textures. A comparison with alternative pre-training methodologies reveals that traditional comparative learning, reliant on positive and negative sample pairs, struggles to discern between positively enhanced variants of the same image and negative samples from different yet visually similar images, leading to convergence challenges. The proposed method adeptly circumvents this limitation, delivering enhanced performance in downstream tasks compared to the pre-training effects observed with large-scale datasets like ImageNet.

This research offers invaluable tools and methodologies for delving into the growth dynamics of cherry plants, skillfully employing techniques like image anomaly detection and curve modeling. Through the application of image anomaly detection, the study efficiently filters out low-quality images, enhancing the reliability and precision of image data analysis. This rigorous selection process ensures the utilization of only high-quality images for further examination, thereby minimizing the influence of errors and noise on the findings. Furthermore, a dataset of high caliber lays a robust groundwork for the subsequent modeling and analysis of growth curves.

This research employs two distinct curve modeling approaches: one based on the pixel proportions of leaves, flowers, and fruits, and the other rooted in the color values of leaves and fruits. These models not only capture the growth trajectory of plant organs with precision but also serve as sensitive indicators of plant health. By integrating Kalman filtering with monotonic processing, the study successfully minimizes data noise and fluctuations, yielding smoother and more dependable curves that provide a solid data foundation for further analysis. The study enhances understanding of the relative changes in plant organs during growth phases. For instance, an increase in the pixel proportion of fruit might signal the onset of fruit development, whereas a declining flower proportion could suggest the fading of flowers. The construction of growth change curves vividly portrays the developmental patterns of various plant organs, offering visual data support for an enriched comprehension of the growth process. This insight is pivotal for discerning the growth patterns and dynamic properties of plant organs, playing a crucial role in grasping the biological behaviors and growth principles of plants. Notably, the use of the Richards curve fitting model, adept at depicting diverse growth trends, is a highlight of this study. For cherry fruits, characterized by two distinct expansion phases often overlooked in conventional growth modeling, this research delineates the fruit growth curve into two segments. This segmentation enables the precise identification of these pivotal growth phases and accurately captures the fruit's coloration period, offering deep insights into fruit development. Such phased growth recognition is invaluable for fruit farmers, allowing for tailored management practices aligned with the fruit's growth stages.

However, it must be acknowledged that there are still certain limitations in the experimental setup of this study. Due to the considerable workload and time required for obtaining time-series images, the experiments in this paper were conducted solely based on cherry time-series images. However, there is a wide variety of plant species, each with different characteristics and growth patterns. Therefore, future research could consider expanding the dataset to include more plant species to improve the model's generalization ability. Additionally, besides image data, there are many other types of time-series data that can be used for

studying plant growth, such as climate data, soil data, etc. These data may contain important information that cannot be captured by image data alone. Hence, in future research, leveraging multimodal data could be considered to enhance the performance of the model.

In conclusion, despite the limitations of the experiments, the analysis method based on PDE still contributes to a profound understanding of the growth characteristics and organ development of cherry plants, providing valuable scientific insights for agricultural management. By meticulously monitoring the development of various organs, this research is capable of pinpointing the ideal harvest time, optimizing the allocation of resources, minimizing waste, and contributing to disease surveillance and ongoing plant health assessment. This plays a crucial role in enhancing agricultural productivity and quality, addressing the escalating global food requirements, and fostering the sustainable advancement of contemporary agriculture.

## 7. Conclusion

This paper proposes a method called self-supervised contrastive learning method for plant time-series images with a Priori Distance Embedding to address the field of knowledge that the semantic information in images corresponding to different phenological periods of plants also varies. Establish pairs of different types of images with distance levels. Building a comparative network based on the Siamese network (Bromley et al., 1993) for comparative learning training.

Migrate to downstream cherry temporal image modeling task. The pre-trained model serves as an encoder, combined with VAE (An and Cho, 2015) for temporal image anomaly detection. Transfer the pre-trained image to the U-Net encoder to segment various organ regions of the plant image. Accumulate data in the time dimension, establish a full cycle growth curve, and further use the richards curve (Richards, 1959) to fit the model, numerically describing the cherry growth process.

This article introduces a novel pre-training paradigm tailored for comparative learning of plant temporal images, employing cherry temporal images as the experimental subject to construct a comprehensive growth model. This serves as an exemplary case of full-process analysis of plant temporal image phenotypes. The self-supervised comparative learning approach presented herein proves to be efficaciously applicable to the pre-training of plant temporal images, heralding wide-ranging potential applications across various computer vision studies related to plant phenotyping.

## CRedit authorship contribution statement

**Wei Xu:** Writing – original draft, Validation, Software, Methodology, Conceptualization. **Ruiya Guo:** Writing – original draft, Methodology. **Pengyu Chen:** Investigation. **Li Li:** Validation, Investigation, Conceptualization. **Maomao Gu:** Investigation. **Hao Sun:** Investigation. **Lingyan Hu:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization. **Zumin Wang:** Funding acquisition, Conceptualization, Project administration, Resources, Supervision. **Kefeng Li:** Funding acquisition, Project administration, Supervision.

## 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.

## Data availability

Data will be made available on request.

## Acknowledgments

We thank the cherry farmers in Dalian, China, for their support of this study. This study was supported by the National Natural Science Foundation of China under Grant 61601076, and Science and Technology Innovation Fund of Dalian under Grant 2020JJ26SN058 and 2021JJ13SN78.

## References

- Aksoy, E.E., Abramov, A., Wörgötter, F., Scharr, H., Fischbach, A., Dellen, B., 2015. Modeling leaf growth of rosette plants using infrared stereo image sequences. *Comput. Electron. Agric.* 110, 78–90.
- An, J., Cho, S., 2015. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture IE 2 (1)*, 1–18.
- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1993. Signature verification using a "siamese" time delay neural network. *Adv. Neural Inf. Process. Syst.* 6.
- Cao, M., Sun, Y., Jiang, X., Li, Z., Xin, Q., 2021. Identifying leaf phenology of deciduous broadleaf forests from phenocam images using a convolutional neural network regression method. *Remote Sens.* 13 (12), 2331.
- Cao, M., Xin, Q., 2021. Vegetation phenology detection of deciduous broad-leaf forest using YOLOv3 from PhenoCam. In: 2021 2nd International Conference on Artificial Intelligence and Education. ICAIE, IEEE, pp. 262–270.
- Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inf. Process. Syst.* 33, 9912–9924.
- Chen, X., Fan, H., Girshick, R., He, K., 2020a. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
- Chen, X., He, K., 2021. Exploring simple siamese representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 15750–15758.
- Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020b. A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning*. PMLR, pp. 1597–1607.
- Chen, Y., Zhang, D., 2022. Integration of knowledge and data in machine learning. *arXiv preprint arXiv:2202.10337*.
- Cui, X., Chen, M., Chen, Z., Xu, F., Wang, X., 2021. Forest phenology recognition method based on attention mechanism. *J. Central South Univ. Forestry Technol.* 41 (07), 11–19.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, pp. 248–255.
- Fchollet, 2020. Keras documentation: Visualizing what convnets learn.
- Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings*, pp. 249–256.
- Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent—a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* 33, 21271–21284.
- Güldenring, R., Nalpantidis, L., 2021. Self-supervised contrastive learning on agricultural images. *Comput. Electron. Agric.* 191, 106510.
- Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2. CVPR'06, IEEE, pp. 1735–1742.
- He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9729–9738.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778.
- Houle, D., Govindaraju, D.R., Omholt, S., 2010. Phenomics: the next challenge. *Nat. Rev. Genet.* 11 (12), 855–866.
- Hu, L., Zhou, T., Xu, W., Wang, Z., Pei, Y., 2022. An improved SqueezeNet lightweight model for tomato disease recognition. *J. Zhengzhou Univ. (Natural Sci. Ed.)* 54 (04), 71–77. <http://dx.doi.org/10.13705/j.issn.1671-6841.2021311>.
- Jaccard, P., 1912. The distribution of the flora in the alpine zone. 1. *New Phytol.* 11 (2), 37–50.
- Kar, S., Nagasubramanian, K., Elango, D., Nair, A., Mueller, D.S., O'Neal, M.E., Singh, A.K., Sarkar, S., Ganapathysubramanian, B., Singh, A., 2022. Self-supervised learning improves agricultural pest classification. In: *AI for Agriculture and Food Systems*.
- Karadavut, U., Palta, Ç., Kökten, K., Bakoğlu, A., 2010. Comparative study on some non-linear growth models for describing leaf growth of maize. *Int. J. Agric. Biol.*
- Kierdorf, J., Junker-Frohn, L.V., Delaney, M., Olave, M.D., Burkart, A., Jaenicke, H., Muller, O., Rascher, U., Roscher, R., 2023. GrowliFlower: An image time-series dataset for growth analysis of cauliflower. *J. Field Robotics* 40 (2), 173–192.

- Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- Kolhar, S., Jagtap, J., 2021. Convolutional neural network based encoder-decoder architectures for semantic segmentation of plants. *Ecol. Inform.* 64, 101373.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2980–2988.
- Margapuri, V., Neilsen, M., 2021. Classification of seeds using domain randomization on self-supervised learning frameworks. In: *2021 IEEE Symposium Series on Computational Intelligence*. SSCI, IEEE, pp. 01–08.
- Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: *2016 Fourth International Conference on 3D Vision*. 3DV, Ieee, pp. 565–571.
- Pan, Y., et al., 2015. Analysis of concepts and categories of plant phenome and phenomics. *Acta Agron. Sinica* 41 (2), 175–186.
- Ranganathan, A., 2004. The levenberg-marquardt algorithm. *Tutorial LM Algorithm* 11 (1), 101–110.
- Richards, F.J., 1959. A flexible growth function for empirical use. *J. Exper. Botany* 10 (2), 290–301.
- Richardson, A.D., Hufkens, K., Milliman, T., Aubrecht, D.M., Chen, M., Gray, J.M., Johnston, M.R., Keenan, T.F., Klosterman, S.T., Kosmala, M., et al., 2018. Tracking vegetation phenology across diverse North American biomes using PhenoCam imagery. *Sci. Data* 5 (1), 1–24.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18. Springer, pp. 234–241.
- Rubinstein, R., 1999. The cross-entropy method for combinatorial and continuous optimization. *Methodol. Computing Appl. Probabil.* 1, 127–190.
- Seyednasrollah, B., Young, A.M., Hufkens, K., Milliman, T., Friedl, M.A., Froelking, S., Richardson, A.D., 2019. Tracking vegetation phenology across diverse biomes using version 2.0 of the PhenoCam dataset. *Sci. Data* 6 (1), 222.
- Song, G., Wu, S., Lee, C.K., Serbin, S.P., Wolfe, B.T., Ng, M.K., Ely, K.S., Bogonovich, M., Wang, J., Lin, Z., et al., 2022. Monitoring leaf phenology in moist tropical forests by applying a superpixel-based deep learning method to time-series images of tree canopies. *ISPRS J. Photogramm. Remote Sens.* 183, 19–33.
- Sun, Z., Li, Q., Jin, S., Song, Y., Xu, S., Wang, X., Cai, J., Zhou, Q., Ge, Y., Zhang, R., et al., 2022. Simultaneous prediction of wheat yield and grain protein content using multitask deep learning from time-series proximal sensing. *Plant Phenom.*
- Tieleman, T., Hinton, G., et al., 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSE: *Neural Netw. Mach. Learn.* 4 (2), 26–31.
- Yasrab, R., Zhang, J., Smyth, P., Pound, M.P., 2021. Predicting plant growth from time-series data using deep learning. *Remote Sens.* 13 (3), 331.