Journal of Field Robotics

WILEY

RESEARCH NOTE

Digital Twin/MARS-CycleGAN: Enhancing Sim-to-Real Crop/Row Detection for MARS Phenotyping Robot Using Synthetic Images

David Liu[1] | Zhengkun Li[2] | Zihao Wu[3] | Changying Li[2]

[1]Department of Computer Science, Princeton University, Princeton, New Jersey, USA | [2]Department of Agricultural and Biological Engineering, University of Florida, Gainesville, Florida, USA | [3]School of Computing, University of Georgia, Athens, Georgia, USA

Correspondence: Changying Li (cli2@ufl.edu)

Received: 11 November 2023 | **Revised:** 18 September 2024 | **Accepted:** 3 November 2024

Keywords: digital twin | field-based robotic phenotyping | object detection | sim-to-real transfer | zero-shot

ABSTRACT

Robotic crop phenotyping has emerged as a key technology for assessing crops' phenotypic traits at scale, which is essential for developing new crop varieties with the aim of increasing productivity and adapting to the changing climate. However, developing and deploying crop phenotyping robots faces many challenges, such as complex and variable crop shapes that complicate robotic object detection, dynamic and unstructured environments that confound robotic control, and real-time computing and managing big data that challenge robotic hardware/software. This work specifically addresses the first challenge by proposing a novel Digital Twin(DT)/MARS-CycleGAN model for image augmentation to improve our Modular Agricultural Robotic System (MARS)'s crop object detection from complex and variable backgrounds. The core idea is that in addition to the cycle consistency losses in the CycleGAN model, we designed and enforced a new DT/MARS loss in the deep learning model to penalize the inconsistency between real crop images captured by MARS and synthesized images generated by DT/MARS-CycleGAN. Therefore, the synthesized crop images closely mimic real images in terms of realism, and they are employed to fine-tune object detectors such as YOLOv8. Extensive experiments demonstrate that the new DT/MARS-CycleGAN framework significantly boosts crop/row detection performance for MARS, contributing to the field of robotic crop phenotyping. We release our code and data to the research community (https://github.com/UGA-BSAIL/DT-MARS-CycleGAN).

1 | Introduction

To sustain the growing world population, crop breeders strive to develop high-yielding and stress-tolerant cultivars that are more resilient to the changing climates, pests, and diseases (Atefi et al. 2021; Rahaman et al. 2015; Jiang and Li 2020). In the past decade, automated robotic crop phenotyping technologies have been developed to assist the assessment and quantification of phenotypic traits related to crop growth, yield, and adaptation to environmental stresses (Iqbal et al. 2020; Xu and Li 2022b; Chawade et al. 2019). One extensively studied, yet still challenging problem in the field of robotic phenotyping is accurate crop detection. Accurately identifying the crop target is a prerequisite for the robot to perform multiple key downstream tasks such as row detection, path planning, navigation, and extraction of phenotypic traits.

Crop phenotyping robots have benefited from the rapid advancements in deep learning and sensor technologies (Qiao et al. 2022; He et al. 2022a). However, the semantic understanding of agricultural images still faces numerous challenges resulting from significant biological variability (e.g., crop species,

David Liu, Zhengkun Li, and Zihao Wu contributed equally to this study.

growth stages, and health conditions) and unstructured environments (e.g., object occlusion, variable lighting conditions, cluttered scenes) (Barth et al. 2020). These challenges highlight the necessity of creating extensive datasets that represent a wide range of variations to develop robust deep learning models. Collecting vast amounts of data and annotating it for diverse crops in different growth stages in the real world is time-consuming and resource-intensive. Currently, the availability of comprehensive, annotated image datasets in agriculture that match the scale of those in general-purpose computer vision is limited (Lu and Young 2020a).

One prominent trend in overcoming the limitations of data set scarcity and variability in agricultural contexts is the utilization of simulation data and simulation-to-reality transfer methods (sim-to-real) (Vierbergen et al. 2023; Katyara et al. 2021). These methods involve synthesizing data through advanced simulation techniques to augment real-world datasets, with the core idea of using digital models of crops and soil grounds to build diverse simulated fields. Similarly, robotic and imaging systems can also be simulated in the virtual world to generate diverse sets of images (Pylianidis et al. 2021), which can enhance the performance of machine learning models by training the models with more diverse data than those available in the real-data sets.

The effectiveness of sim-to-real transfer hinges critically on the realism of the synthetic data. Therefore, the limitations of this approach are primarily tied to the fidelity of the simulation models used, including the accuracy of crop models, the realism of texture rendering, and the sophistication of environmental lighting and shading effects. Additionally, the variations in camera perspectives, angles, and views in simulations versus real-world conditions can introduce discrepancies that may affect the performance of the trained models when applied in actual field settings (Truong et al. 2023; Höfer et al. 2020). The generative adversarial networks (GANs) (Goodfellow et al. 2020) have been proven to be an effective method for synthesizing large-scale realistic images for model training to deal with the reality gap problem in sim-to-real transfer (Kleeberger et al. 2020; Bousmalis et al. 2017; Lu and Young 2020a; Peng et al. 2018; Patel et al. 2015). For instance, the retinaGAN (Ho et al. 2021) and RL-CycleGAN (Rao et al. 2020) were successfully applied to enhance sim-to-real robotic grasping.

Along this promising research direction and inspired by the recent research in (Liu et al. 2023), this paper specifically proposes a novel DT/MARS-CycleGAN model for enhancing sim-to-real crop/row detection in the context of robotic phenotyping with our customized Modular Agricultural Robotic System (MARS) (Xu and Li 2022a; Li et al. 2022). Our contributions are as follows: (1) Generating diverse, auto-labeled simulation images to mitigate the scarcity of labeled real images using digital twin (DT) of MARS robots: real and virtual DT MARS robots are forced to mimic each other such that the gaps between simulated and realistic robotic crop phenotyping are minimized at a fundamental level. (2) Inspired by previous work (Liu et al. 2023), we designed a new DT/MARS loss tailored for precision agriculture. This loss is integrated into the CycleGAN model (Zhu et al. 2017) to enforce object position consistency and further bridge the reality gap between synthetic images generated by the DT MARS and real crop images captured by the physical MARS. (3) Sim-to-real crop/row detection: we validate the effectiveness of crop/row detection on real scenarios using the fine-tuned crop detectors with the synthesized images. Compared to other simulation or GAN-based methods, the synthetic crop images generated by DT/MARS-CycleGAN exhibit significant realism, closely resembling real images. As a result, they provide a more effective foundation for fine-tuning object detectors such as YOLOv8 (Solawetz 2023), enhancing their performance in real-world scenarios. Extensive experimental results have demonstrated that DT/MARS-CycleGAN framework significantly narrowed the gap between the simulation and reality and improved the successful rate of sim-to-real transfer, thus contributing to robotic crop phenotyping and precision agriculture.

In this paper, we present DT/MARS-CycleGAN, leveraging the diverse, auto-labeled simulation images to enhance crop/row detection in precision agriculture. The proposed model applies consistency loss not only on image appearance but also on object position, achieving zero-shot sim-to-real image synthesis to mitigate the scarcity of labeled real images in real-world scenarios. The rest of this paper is organized as follows. In Section 2, we briefly review the state-of-the-art in robotic phenotyping, digital twin, and Generative Adversarial Networks in precision agriculture and phenomics. Then the details of DT/MARS-CycleGAN framework are introduced in Section 3 and evaluations of its sim-to-real ability in crop and row detection, compared with other classic GAN-based methods in Section 4. Finally, the main findings, limitations, and future directions are discussed and concluded in Section 5.

2 | Related Works

2.1 | Crop/Row Detection in Robotic Phenotyping

Single crop or row (Crop/row) detection is critical for robotic phenotyping in agriculture, enabling automated systems to navigate fields and perform tasks such as measuring crop traits with high efficiency and minimal human intervention (Xu and Li 2022b; He et al. 2022b). Initially, detection algorithms primarily utilized image features such as crop row color and texture (Wang et al. 2022), which are highly susceptible to variations in imaging conditions. These early techniques, including color-index-based and threshold-based segmentation, faced significant challenges under varying lighting and environmental conditions (Bai et al. 2023).

The adoption of deep learning has marked a significant advancement in precision agriculture, as seen in the DeepFruit project, which utilized the Faster RCNN model (Sa et al. 2016; Girshick et al. 2014). Recent research related to row detection has achieved significant performance optimizing the crop row extraction through combining line extraction algorithms such as HT (Hough transform), LR (linear regression), and HF (horizontal fringes) (Huang et al. 2021; De Silva et al. 2024; Ahmadi, Halstead, and McCool 2021; Winterhalter et al. 2021; Liang et al. 2022). These techniques have shown superior

performance in various contexts, overcoming some limitations of traditional methods (Zhang et al. 2024).

However, deep learning applications in agriculture face unique challenges, particularly in data collection and handling intra-field variability (Tian et al. 2020). Recent datasets such as LincolnBeet (Salazar-Gomez et al. 2022), Sugarbeet2016 (Chebrolu et al. 2017), and PhenoBench (Weyler et al. 2024) provide valuable resources but often lack the diversity needed to train models that generalize well across different agricultural settings (e.g., plant development, weather fluctuations, and variable lighting conditions) (Lu and Young 2020b). The complexity of agricultural environments and the high cost of data collection pose significant barriers, although emerging solutions such as synthetic data generation and transfer learning are beginning to address these issues (Polvara et al. 2024).

2.2 | Digital Twin in Agriculture Robotics

The concept of "digital twins" was described as a virtual, digital equivalent (representation) of a physical product and the bidirectional flow of data between them (Jones et al. 2020). A DT in agriculture encompasses sensors, IoT devices, data analytics, and machine learning models, which collectively create a dynamic digital replica of farming operations (Liu et al. 2021; Purcell and Neubauer 2023). The idea of virtual representation via DT technologies significantly enables and facilitates the digitization of agriculture, in which data, modeling, and what-if simulations are integrated to provide a promising framework to overcome constraints in decision-making support and automation for various agricultural applications, including plant monitoring (Moghadam et al. 2020; Angin et al. 2020), soil management (e.g., irrigation and fertilization) (Jayaraman et al. 2016; Alves et al. 2019) and equipment optimization (Kampker et al. 2019).

Benefiting from the powerful physical simulators (e.g., Unreal, Gazebo, PyBullet, and NVIDIA Issac), DT technologies have been explored to realize the sim-to-real ability that enables learning or training robots in the digital/simulation world and deploying robots into real physical scenarios (Huang et al. 2021; Liu et al. 2021; Liu 2022). For instance, researchers reported a DT-enabled approach for achieving effective transfer of deep reinforcement learning (DRL) algorithms to a physical robot (Liu 2022). In another study, the authors acquired large-scale visual grasping datasets with ground truth annotations in a DT environment that mimics the real world to train their DT-CycleGAN model and reported promising results for zero-shot sim-to-real transfer of grasping models (Liu et al. 2023). In this work, we focus on the sim-to-real direction in the bidirectional DT flow rather than real-to-sim simulation, which involves real-time behavior synchronization based on sensors. This focus is driven by our main goal: to generate realistic data to mitigate the high cost of manual data collection.

In general, DT technology has mitigated the scarcity of labeled data in precision agriculture and robotics at a systemic level and offers more precise and controllable data customization, thereby enriching the diversity of image structure, background, and target objects. However, the simulation images generated by DT rely on the fidelity and variety of 3D models, which creates certain gaps in representing real crop fields. Instead, the proposed DT/MARS-CycleGAN model takes advantage of integrating DT with GANs in both agriculture and robotics by maximizing the similarity between the physical and virtual DT MARS robots so that their perceptions align, minimizing the reality gap in zero-shot sim-to-real transfer.

2.3 | Generative Adversarial Networks in Agriculture

The generative adversarial networks (GANs) model (Goodfellow et al. 2020) and their variants (Zhu et al. 2017; Brock 2018; Karras et al. 2020; Esser, Rombach, and Ommer 2021) have been widely used to generate synthesized image data for model training (Khalifa et al. 2022; Lu et al. 2022), including in precision agriculture applications such as multi-species classification (Madsen, Dyrmann, et al. 2019; Madsen, Mortensen, et al. 2019), plant vigor assessments (Zhu et al. 2020; Drees et al. 2021), pest/disease detection (Karam et al. 2022; Bi and Hu 2020), and crop yield estimation (Shete et al. 2020; Hartley and French 2021).

However, synthesized images by GANs still present the challenges of achieving realistic diversity and stability. Several studies focused on improving the generation of realistic images to deal with the reality gap problem in sim-to-real transfer in the general robotics domain (Kleeberger et al. 2020). For example, a novel DT-CycleGAN (Liu et al. 2023) was proposed to minimize the reality gap effectively by integrating two mainstream image data augmentation methods of the generative CycleGAN model and simulations in the DT space. This DT-CycleGAN model inspired us to conduct simulations to generate synthesized crop images in the virtual space and employ a novel DT-MARS consistency loss to make the synthesized crop images more similar to real-world ones captured by our MARS phenotyping robot.

3 | Methods

As illustrated in Figure 1, our framework consists of three main components: the robots operational in both real-world and simulated environments, the crop object/row detection network, and the DT/MARS-CycleGAN model that integrates the robot and the object detection network.

3.1 | Physical and Digital-Twin Robots

Figure 1 (left panel) showcases the MARS phenotyping robot (PhenoBot) (Li et al. 2022) employed in this work. The MARS PhenoBot is a solar-powered modular platform with a four-wheel steering and four-wheel driving configuration. Each wheel module of MARS PhenoBot is equipped with an independent suspension mechanism, which makes it adaptable to uneven field terrain. The MARS PhenoBot robot is designed for streamlined modular phenotyping and specifically tailored for crop phenotyping (Xu and Li 2022a; Li et al. 2022). The robot is outfitted with three cameras, capturing views from the front,

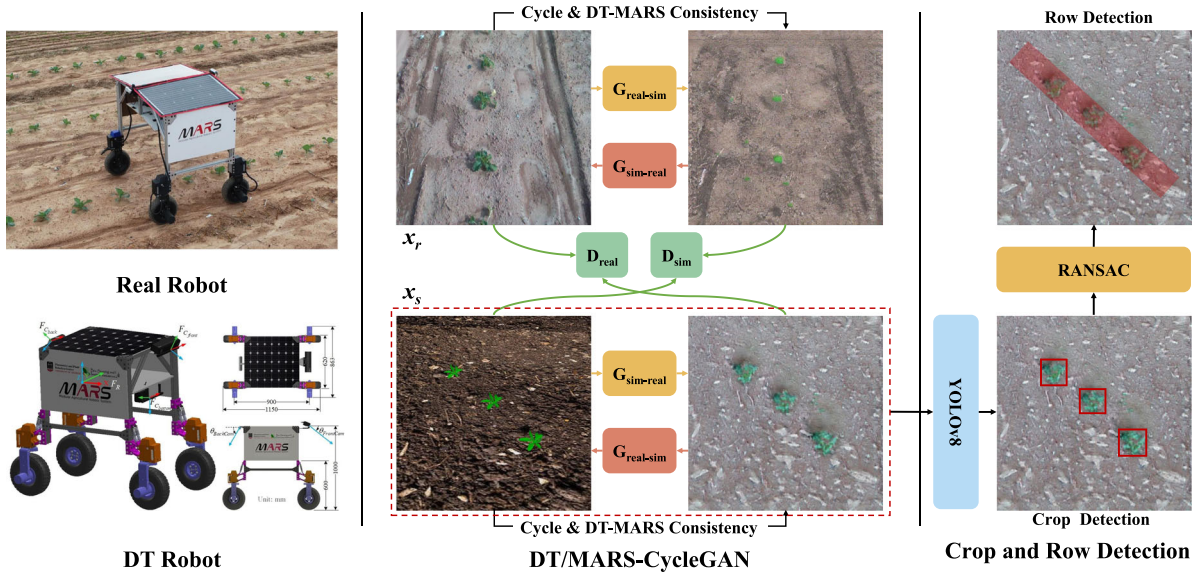

FIGURE 1 | Illustration of the physical/DT robots (left panel), the DT/MARS-CycleGAN model (middle panel), and the crop object/row detection network (right panel). Additional annotations and descriptions are explained in the main text. [Color figure can be viewed at wileyonlinelibrary.com]

rear, and bottom, respectively. This setup allows for effective monitoring of crop growth and offers crucial row-specific data to enhance farm field navigation and to prevent crop damage. The bottom camera is specifically oriented toward the ground to observe more details of the farm field and to detect crops and rows. More details about the MARS PhenoBot can be found in Li et al. (2022). The virtual DT robot of MARS PhenoBot was developed using the Solidworks tool (Dassault Systèmes, Waltham, USA). Special attention was given to simulating the camera perspectives (e.g., bottom camera for crop/row detection) to reduce discrepancies in visual sensing between the real robot and the DT version, thereby narrowing the reality gap during zero-shot sim-to-real transfer of crop object detection models.

3.2 | Crop Object/Row Detection Network

The crop object/row detection network consists of two steps: crop detection is carried out using the YOLOv8 object detection model (Solawetz 2023), followed by the RANdom SAmple Consensus (RANSAC) algorithm (Fischler and Bolles 1981) to delineate crop rows.

As depicted in Figure 1, crop field images, along with their corresponding crop object annotations, are initially fed into the YOLOv8 model. The input image is represented as $\mathbf{x} \in \mathbb{R}^{B\times C\times H\times W}$, where $B \times C \times H \times W$ denotes the batch size (B), channel (C), height (H), and width (W) of the images. For expedited training, images are resized to 224×224. The bounding box annotation is formatted with one row per object represented as $\mathbf{y} \in \mathbb{R}^{Bn4}$, where B is the batch size, n represents the number of objects in an image, each described in $[x_{center}, y_{center}, width, height]$ format. By optimizing the IoU loss between the predicted bounding boxes and the ground truth, the YOLOv8 model can accurately locate the crop objects in the images.

With the crop bounding boxes predicted by YOLOv8, the centers of all bounding boxes are extracted together to form the point data essential for row detection. Subsequently, the RANSAC algorithm is employed to discern a line representing the crop row that best fits the center points of the bounding boxes of crop objects. The detected crop row is then used as the input for visual servoing control of the MARS PhenoBot. That is, the error offset of the detected crop row from the robot's reference is used as the signal input to a PID (Proportional-Integral-Derivative) controller to control the robot's velocity and movement (Li et al. 2022). It is apparent that accurate detection of crops and crop rows is a fundamentally important task for crop phenotyping robot's path planning and navigation.

For zero-shot sim-to-real transfer of the crop detection model trained in a simulated environment, the object detection model along with both the images and the bounding box predictions, is holistically modeled by the DT/MARS-CycleGAN model.

3.3 | DT/MARS-CycleGAN: Cycle Consistency Loss

The CycleGAN introduced by Zhu et al. (2017) was originally developed for unpaired image style transfer. In our work, it is adapted and integrated into the DT/MARS-CycleGAN model to address the discrepancies in visual appearance between simulated and real-world images, as demonstrated by the real-to-sim and sim-to-real transitions in the middle panel of Figure 1. Here, $\mathbf{x_r} \in \mathbf{R}$ represents real images, and $\mathbf{x_s} \in \mathbf{S}$ denotes simulated images, as illustrated in Figure 1. The simulated images are acquired from the Pybullet[1] simulation environment, created by utilizing farm field images collected from the Internet as background and rendering 3D models of the crops as target objects at random positions. A virtual camera mounted on the DT robot captures these images within the simulated environment, and an analytical virtual model extracts the ground truth

bounding boxes for each crop in the image based on the predefined position information. Meanwhile, real images were obtained from actual farm environments by capturing video streams as the real robot traversed the crop rows in the farm field.

Specifically, as delineated in the DT/MARS-CycleGAN panel in Figure 1, the generator, $G_r(\cdot)$, transforms images to a real style, and $G_s(\cdot)$ modifies the input images to a simulation style. These correspond to the generators $F(*)$ and $G(*)$ in Zhu et al's original CycleGAN model (Zhu et al. 2017). The discriminator, $D_r(\cdot)$, classifies whether the input images are in real style, while $D_s(\cdot)$ discerns if the images exhibit a simulation style. Notably, all losses are computed based on $\mathbf{x_r} \in \mathbf{R}$ and/or $\mathbf{x_s} \in \mathbf{S}$. While these parameters are omitted in the following equations for simplicity, they are implicitly involved in all loss calculations. The cycle loss function of DT/MARS-CycleGAN is mathematically described as:

$$L(G_s, G_r, D_s, D_r) = L_{GAN}(G_s, D_s) + L_{GAN}(G_r, D_r) + L_{cyc}(G_r, G_s) + L_{identity}(G_r, G_s). \quad (1)$$

Here, $L_{GAN}(G_s, D_s)$ and $L_{GAN}(G_r, D_r)$ denote the adversarial losses; $L_{cyc}(G_r, G_s)$ signifies the cycle consistency loss, and $L_{identity}(G_r, G_s)$ represents the identity mapping loss.

Adversarial losses aim to align the style of the generated crop images with the style of the target domain, which is either a real-world farm environment or a simulation environment. Specifically, the generators G and the discriminators D are adversaries in a game during training, where the G learns to produce increasingly realistic data, while the D gets better at distinguishing real data from simulated data. In the following equations, we use $\mathbb{E}\mathbf{x} \sim p_{data}()$ to represent the expected value over the data distribution within a specific domain (i.e., $\mathbf{S}$ or $\mathbf{R}$). The adversarial losses are expressed as follows:

$$L_{GAN}(G_s, D_s) = \mathbb{E}\mathbf{x_s} \sim p_{data}(\mathbf{S})[logD_s(\mathbf{x_s})] + \mathbb{E}\mathbf{x_r} \sim p_{data}(\mathbf{R})[log(1 - D_s(G_s(\mathbf{x_r})))], \quad (2)$$

$$L_{GAN}(G_r, D_r) = \mathbb{E}\mathbf{x_r} \sim p_{data}(\mathbf{R})[logD_r(\mathbf{x_r})] + \mathbb{E}\mathbf{x_s} \sim p_{data}(\mathbf{S})[log(1 - D_r(G_r(\mathbf{x_s})))]. \quad (3)$$

To inhibit the mapping of the input image to any random permutation of images in the target domain style, the cycle consistency loss is utilized, ensuring that the images can be accurately reverted to their original domain after undergoing transformations by two generators in sequence. Here the notation $\|\cdot\|_1$ denotes the L1 norm, which measures the difference between the original and transformed images. The loss is defined as:

$$L_{cyc}(G_r, G_s) = \mathbb{E}\mathbf{x_s} \sim p_{data}(\mathbf{S})[\|G_s(G_r(\mathbf{x_s})) - \mathbf{x_s}\|_1] + \mathbb{E}\mathbf{x_r} \sim p_{data}(\mathbf{R})[\|G_r(G_s(\mathbf{x_r})) - \mathbf{x_r}\|_1]. \quad (4)$$

Lastly, the identity mapping loss preserves the content consistency of the image, formulated as:

$$L_{identity}(G_r, G_s) = \mathbb{E}\mathbf{x_s} \sim p_{data}(\mathbf{S})[\|G_s(\mathbf{x_s}) - \mathbf{x_s}\|_1] + \mathbb{E}\mathbf{x_r} \sim p_{data}(\mathbf{R})[\|G_r(\mathbf{x_r}) - \mathbf{x_r}\|_1]. \quad (5)$$

3.4 | DT/MARS-CycleGAN: DT-MARS Consistency Loss

Inspired by the DT-CycleGAN model originally designed for single-object visual grasping tasks (Liu et al. 2023), we introduce a new DT-MARS consistency loss tailored for crop detection in precision agriculture. This adaptation ensures more coherent sim-to-real transfer by penalizing positional and size shifts across all detected crops, addressing the complexities of outdoor agricultural environments.

Specifically, given a simulation image $\mathbf{x} \in \mathbf{S}$, the position and size of the crops, represented by the predicted bounding boxes $\mathbf{b}_1 = \text{Detector}(\mathbf{x})$, should align with those of the sim-to-real transferred image, $\mathbf{b}_2 = \text{Detector}(G_r(\mathbf{x}))$. Unlike previous work, which focused on single-object detection, our DT-MARS loss operates across all objects to minimize overall positional shifts, enabling accurate multi-crop detection. As shown in the middle panel of Figure 1, the loss uses the L1 norm to measure the difference between the bounding boxes of the original and generated images, and is formulated as follows:

$$\begin{aligned} &L_{DT-MARS}(G_r, G_s, \text{Detector}) \\ &= \mathbb{E}\mathbf{x_s} \sim p_{data}(\mathbf{S})[\|\text{Detector}(G_r(\mathbf{x_s})) - \text{Detector}(\mathbf{x_s})\|_1] \\ &+ \mathbb{E}\mathbf{x_r} \sim p_{data}(\mathbf{R})[\|\text{Detector}(G_s(\mathbf{x_r})) - \text{Detector}(\mathbf{x_r})\|_1]. \end{aligned} \quad (6)$$

Thus, the total loss function for DT/MARS-CycleGAN model is represented as the sum of the cycle consistency loss and the DT-MARS spatial consistency loss:

$$\begin{aligned} &L(G_s, G_r, D_s, D_r, \text{Detector}) \\ &= \lambda_{\text{gan}} \times L_{\text{GAN}}(G_s, D_s) + \lambda_{\text{gan}} \times L_{\text{GAN}}(G_r, D_r) \\ &+ \lambda_{\text{cyc}} \times L_{\text{cyc}}(G_r, G_s) + \lambda_{\text{identity}} \times L_{\text{identity}}(G_r, G_s) \\ &+ \lambda_{\text{detector}} * L_{\text{DT-MARS}}(G_r, G_s, \text{Detector}). \end{aligned} \quad (7)$$

In this expression, $\lambda_{\text{gan}}, \lambda_{\text{cyc}}, \lambda_{\text{identity}}$, and $\lambda_{\text{detector}}$ are the hyperparameters governing the impact of the respective loss components in the overall loss function. In the model training and evaluation process, the values experimentally assigned to these hyperparameters are $\lambda_{\text{gan}} = 1, \lambda_{\text{cyc}} = 5, \lambda_{\text{identity}} = 2$, and $\lambda_{\text{detector}} = 10$. This setting of hyperparameters was pre-tuned and inherited from the previous work (Liu et al. 2023), aiming to balance the contributions of each loss component and prevent any single component from dominating. This ensures the generation of images that are both realistic and content-consistent. The setting maintains a dynamic balance between each loss component throughout the training process and demonstrates potential robustness with minimal tuning across different applications and scenarios.

The introduction of the DT-MARS consistency loss ensures that the bounding boxes, representing the DT crops in both real and simulated domains, are consistent in terms of their positions and dimensions, leading to more robust and accurate sim-to-

real transfer by mitigating the discrepancies between these two domains.

3.5 | Training Strategies

The training process of the complete framework was divided into two stages: the first step involved the training of the crop detection model, and the second engaged the DT/MARS-CycleGAN for zero-shot sim-to-real transfer. Both training stages were conducted exclusively within a simulated environment, without the need for human annotation on the real-world farm images.

3.5.1 | Training Crop Detection Model

Initially, we constructed the simulation environment by utilizing the Pybullet package. The real farm images were utilized as the background for the simulation environment, and placed 3D models of crops were used as target objects at random positions. To enrich the diversity of the generated simulation images, we collected 500 distinct, publicly available images of farm fields and created 3D models of three different crop species: sweet beet, polygonum, and Cirsium at varying growth stages. Subsequently, images were captured by utilizing a virtual camera positioned on the DT robot (bottom view for crop/row sensing) within the simulation environment, and an analytical virtual model extracted the ground truth bounding boxes for each crop in the image based on the predefined position information.

The generated simulation crop detection data set with mimicked appearances and accurate bounding box annotations was then employed for crop detection model training. Given the hardware constraints of real crop robots and the demand for real-time processing in robotic phenotyping tasks, the YOLOv8n detection model (Solawetz 2023) was used for lightweight deployment and swift inference. Specifically, the YOLOv8n weights pre-trained on the COCO data set served (Lin et al. 2014) as the initialization, followed by a fine-tuning step on the generated simulation crop detection data set for domain adaptation.

3.5.2 | Training DT/MARS-CycleGAN

By incorporating unlabeled real image frames extracted from video streams captured by the real robot that navigated along crop rows in actual farm environments, the reality gap in zero-shot sim-to-real transfer was notably reduced during the training of DT/MARS-CycleGAN within the simulation environment. Specifically, 2400 simulation images with automatically annotated bounding boxes and 2400 real frames extracted from the video stream were employed. The training configurations aligned with those described in (Zhu et al. 2017). In contrast to RetinaGAN (Ho et al. 2021), where the sim-to-real transfer process solely functions as a method to augment data for training the detection model, our framework facilitated the simultaneous training of the crop detection model and the DT/MARS-CycleGAN. Subsequent experiments revealed that this novel methodology achieves superior results when compared to the RetinaGAN model (Ho et al. 2021).

4 | Evaluation

4.1 | Data Set and Evaluation Metrics

4.1.1 | Data Set

The training set comprised of images from two domains: 2400 auto-annotated simulation crop detection data generated from the PyBullet simulation environment, and 2400 real frames extracted from the video stream as previously detailed. To test the performance of the trained model, we assembled a high-quality testing set including 408 unseen image frames gathered by the real robot in farm fields. Each image was manually labeled by utilizing the Roboflow labeling tool (Roboflow, Inc., Des Moines, USA), without data augmentation, and was formatted in the YOLO object detection data set format. Statistics of the number of images of different species crops in our collected simulated and real crop datasets were shown in Table 1.

In addition, an open-source sugar beet data set, the Lincolnbeet data set (Salazar-Gomez et al. 2022) was used as an additional test set for validating the performance of transfer learning. The Lincolnbeet data set was an object detection data set designed to encourage research in the identification of items in environments with high levels of occlusion, and in the development of better approaches to evaluate object detection models in practical scenarios. These images were collected on four different dates a week apart (May–June 2021) to record crops/weeds at different growth stages with a resolution of 1920×1080. A total of 4402 images containing 38,234 crop instances were divided into training, validation and testing sets with a ratio of 7:2:1.

In summary, we trained the detection models on simulation and generative images, and tested them on manually collected in-field images set and the previously unseen Lincolnbeet testing set. To further assess the model's effectiveness in real-world

TABLE 1 | Statistics of the simulation and real data set. Collard and cabbage are grouped due to their similarity.

Domain	Train	Test
Sim	800 Seedling Sugar beet	—
	800 Well-grown Sugar beet	—
	400 Polygonum	—
	400 Cirsium	—
Real	~1500 Collard and Cabbage	341 Collard and Cabbage
	~900 Kale	67 Kale
Total	2400 sim for III-E.1	408 real for III-E.1
	2400 real for III-E.2	

scenarios, we also conducted downstream row detection experiments in real fields.

4.1.2 | Evaluation Metrics

Fréchet Inception Distance (FID) and Inception Score (IS) are employed to quantitatively evaluate the appearance fidelity of the sim-to-real transferred images to real ones.

- **FID:** Measures the similarity between the distribution of generated images and the distribution of real image; lower is better. To compute Fréchet Inception Distance, we pass generated and true data through an ImageNet pre-trained Inception V3 model to obtain visually relevant features. Let (M_t, C_t) and (M_g, C_g) represent the mean and covariance of the true and generated features respectively, then compute (Xu et al. 2018):

$$FID = \|M_t - M_g\|_2^2 + \mathrm{Tr}(C_t + C_g - 2(C_t C_g)^{1/2}), \tag{8}$$

where Tr denotes the trace of the matrix, representing the sum of its diagonal elements.

- **IS:** Evaluates the diversity and quality of generated image; high scores indicate that the model generates diverse, high-quality images that are similar to the real images. Let G denote an image generator to be evaluated, and x is an image generated by G. Define $p(y|x)$ as the predicted class distribution from Inception V3 model on image x, and $p(y)$ as the marginal distribution of the predicted labels across all generated images. $D_{KL}(p\|q)$ denotes the KL-divergence between two probability distributions p and q. The Inception Score for G is given by (Salimans et al. 2016):

$$IS(G) = \exp(\mathbb{E}\mathbf{x} \sim p_{data}[D_{KL}(p(y|x)\|p(y))]). \tag{9}$$

Specifically, the evaluated generator $G = G_r$, which transforms the image to the real style, and the generated image $x = G_r(x_s)$, where x_s is the simulated image.

In assessing the detection performance of YOLOv8 trained on sim-to-real transferred images, we adopted common Precision, Recall, F1, mAP50, and mAP50-95 as evaluation metrics (Lin et al. 2014). The following parameters were used in the formulae for some of the above evaluation metrics: TP (predicted as a positive sample and actually as a positive sample as well), FP (predicted as a positive sample, though it is actually a negative sample), and FN (predicted as a negative sample, though it is actually a positive sample). Intersection over Union (IoU) represents the ratio of intersection and concatenation between the bounding box and the true box.

- **P (Precision):** Precision is the ratio of the number of positive samples predicted by the model to the number of all detected samples.

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{10}$$

- **R (Recall):** Recall is the ratio of the number of positive samples correctly predicted by the model to the number of positive samples that actually appeared.

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{11}$$

- **F1 Score):** F1 is the harmonic mean of precision and recall, providing a balanced measure that accounts for both metrics.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{12}$$

- **AP (Average Precision):** The average precision is equal to the area under the precision-recall curve.

$$AP = \int_0^1 \text{Precision}(\text{Recall})\, d(\text{Recall}). \tag{13}$$

This integral evaluates the precision at every increment of recall (101-point interpolation method), from 0 to 1, providing a single-figure summary of model performance across all levels of recall.

- **Mean Average Precision (mAP):** mAP is the average of the AP calculated for all categories, providing a single metric to summarize the performance across multiple classes:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i, \tag{14}$$

where AP_i is the AP for category i, and N is the total number of categories. Specifically, in our data set with only one category (crops), N would be 1. Specific variations of $mAP50$ and $mAP50 - 95$ focus on the model performance at specific IoU (Intersection over Union) thresholds:

- **mAP50:** calculates mAP at an IoU threshold of 0.5, commonly used for evaluating the detection performance at a moderate level of localization accuracy.
- **mAP50-95:** averages mAP calculated at IoU thresholds from 0.5 to 0.95 in increments of 0.05, providing a comprehensive evaluation over a range of strictness in localization accuracy.

4.2 | Zero-Shot Sim-to-Real Transfer

Four distinct zero-shot sim-to-real transfer methods were assessed on the created data set:

- Sim-Only: Training solely on simulation data, without any sim-to-real transfer.
- CycleGAN: Training on sim-to-real transferred images generated by CycleGAN.
- RetinaGAN: Training on sim-to-real transferred images with additional perception consistency, generated by RetinaGAN.
- The proposed DT/MARS-CycleGAN.

YOLOv8n was chosen as a crop detection model for lightweight deployment and the necessity of real-time inference in real-world agricultural applications. In all four sim-to-real transfer methods, the YOLOv8n detection model utilized the pre-trained weights on the COCO data set for initialization and is trained

for 100 epochs with a batch size of 64, learning rate of 1e−2, weight decay of 5e−4, a cosine learning rate scheduler and Adam optimizer. In YOLOv8n, the decoupled head structure used two separate branches to realize object classification and localization. For the classification task, binary cross-entropy loss (BCE loss) was used. For the bounding box regression task, distribution focal loss (DFL) (Li et al. 2020) and CIoU (Zheng et al. 2020) were employed. To ensure zero-shot sim-to-real transfer, all four models were exclusively trained in simulation environments without any fine-tuning on labeled real-world data.

Figure 2 displays the visualizations of generated sim-to-real images by different methods. Although CycleGAN may effectively transfer background information within an image, spatial information pertinent to the crop target is often distorted or even lost. With the additional consistency loss of target detection, RetinaGAN can somewhat adequately preserve target information; however, the background transformed by RetinaGAN appears uniform and repetitive, lacking in variety. This could be due to the fixed detection model parameters in RetinaGAN, which constrain the model within the simulation domain and hamper its adaptability in the real domain, leading to an incomplete transfer of background style. In contrast, DT/MARS-CycleGAN exhibits superior sim-to-real transfer consistency in both appearance and spatial dimensions by enabling the concurrent training of the crop detection model and the DT/MARS-CycleGAN model. This emphasizes the effectiveness of our framework in addressing reality gaps in zero-shot sim-to-real transfer scenarios.

For quantitative analysis, as shown in Table 2, We reported the FID and IS in the Appearance Fidelity panel to assess the fidelity of the transferred images to real ones. In the Crop Detection panel, the detection results of YOLOv8 were reported, where YOLOv8 was trained on images transferred by different models and then tested on the same real images. The proposed DT/MARS-CycleGAN achieved the best performance on both parts, demonstrating promising zero-shot sim-to-real transfer capability. In contrast, due to the target position shift and monotonous background appearance, both the CycleGAN and RetinaGAN methods yielded even poorer performance than Sim-Only. Qualitative visualizations of crop detection for these methods were presented in Figure 3. In the top three rows, the proposed DT/MARS-CycleGAN model significantly reduced both false positives and false negatives compared to other methods. The last row showed a false negative case of the DT/MARS-CycleGAN model, where the undetected object was partially visible at the bottom edge of the camera view under low light conditions.

4.3 | Ablation Studies on Simulation Image Generation

To assess our simulation image generation, we conducted ablation studies from three aspects as follows. All experiments were trained on pure simulation images with different generation setups and tested on the same set of real images used in Table 2.

4.3.1 | Evaluation on Different Training Image Quantity

The proposed DT/MARS-CycleGAN model significantly reduced the human effort required to accumulate labeled real-world training data with crop object bounding boxes. By eliminating this laborious endeavor, substantial amounts of simulation data can be harnessed to enhance zero-shot sim-to-real transfer further. To quantify the influence of the size of simulation data, we trained the crop detection model with varying sample sizes in the simulation environment. As depicted in Figure 4, the mAP50 elevated from 0.903 to 0.917 by increasing

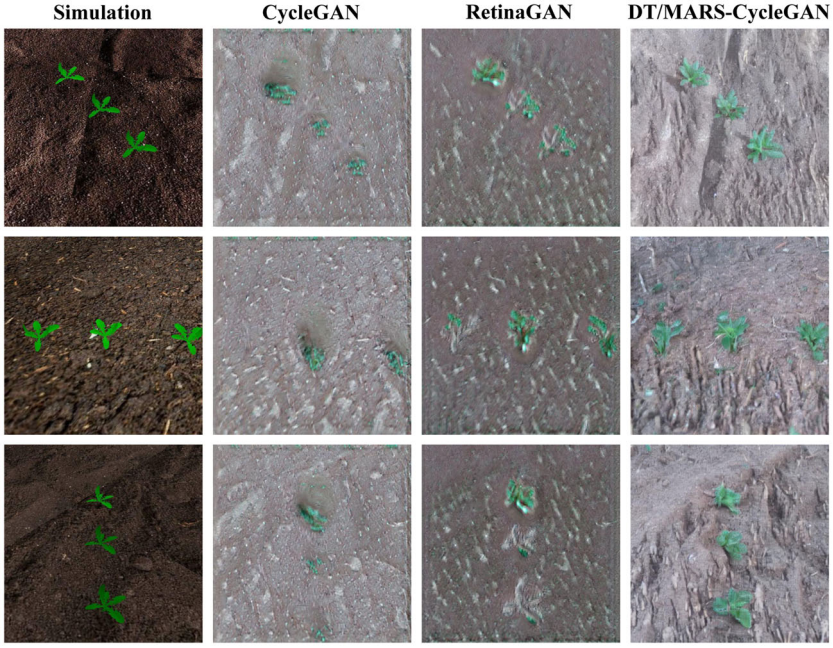

FIGURE 2 | Simulation-to-real synthetic images generated by different models. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 | Comparison of sim-to-real transformation quality across different methods. Appearance fidelity is evaluated using the FrÉchet Inception Distance (FID) and Inception Score (IS), while Crop Detection with YOLOv8 is assessed using typical metrics (Precision-P, Recall-R, mAP50, and mAP50-90).

	Appearance fidelity		Crop detection			
Method	**FID ↓**	**IS ↑**	**P**	**R**	**mAP50**	**mAP50-95**
Sim-Only	—	—	0.925	0.827	0.917	0.656
CycleGAN	7.318	2.296	0.500	0.422	0.411	0.156
RetinaGAN	7.496	2.581	0.741	0.573	0.650	0.221
DT/MARS	**6.562**	**2.825**	0.942	0.895	**0.964**	**0.674**

Note: The bold values indicate the best results among the compared methods for each evaluation metric.

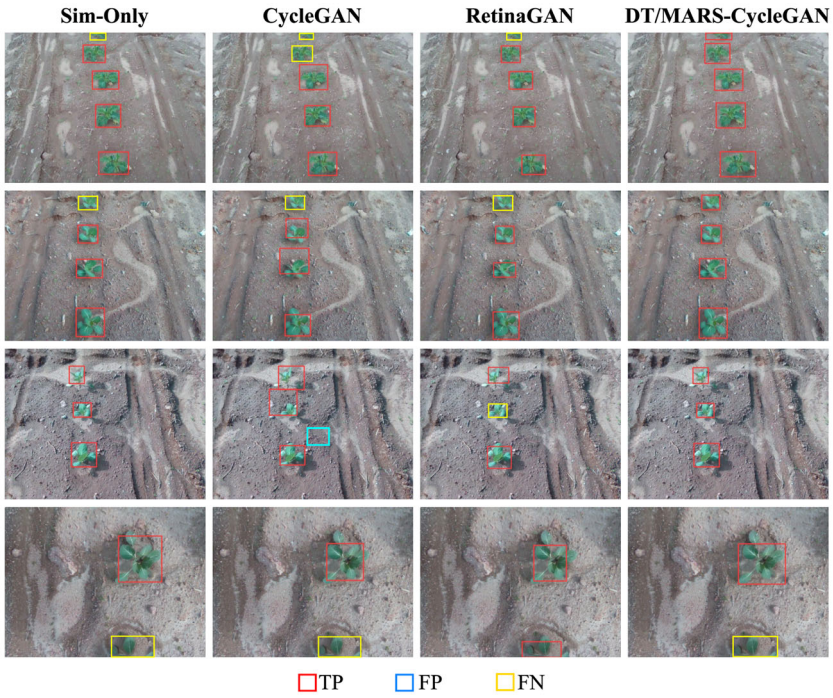

FIGURE 3 | Visualization of crop detection in real images from trained models using different training data (simulation-only, CycleGAN, RetinaGAN, and DT/MARS-CycleGAN). The last row shows a false negative case of the DT/MARS-CycleGAN model. [Color figure can be viewed at wileyonlinelibrary.com]

simulation data from 100 to 2400, without adding any real data requiring costly human annotation. The mAP50-90 similarly ascended from 0.607 to 0.656. Furthermore, the precision and recall were also balanced, resulting in an improved F1 score shown in the figure. It was noteworthy that augmenting simulation data is not only effortless but also enriches the diversity of the training data set, as the simulation images vary in background, object classes, and positions. These findings emphasized that the proposed DT/MARS-CycleGAN has very promising potential in achieving higher-quality and spatial consistency sim-to-real transfer by simply scaling up the synthetic training data from large-scale, diverse simulation data.

4.3.2 | Evaluation on Different Crop Species and Growth Stages

Crop species and growth stages can significantly influence a model's generalizability, as appearances across different species and growth stages can be either highly similar or distinctly different. Selecting an optimal set of crop species and growth stages for generating simulation images can improve the performance of crop detection models pre-trained on such simulation data. In Table 3, the detection models were trained on simulation images featuring various species. The results showed that sugar beet achieves better detection performance compared to other species when used alone for training. When using a combination of two species for training, there was a significant drop in performance if sugar beet was excluded. This was likely because its appearance closely resembles that of collard and kale in the test images. Additionally, training data that combines all species outperformed data from any single species, with two-species combinations partially improving over single-species training. This suggested that diverse species combinations enrich training data and enhance model generalizability. Table 4 presented the detection performance based on training with different growth stages. Similarly, mixed data covering all growth stages surpassed training that focuses solely on one stage.

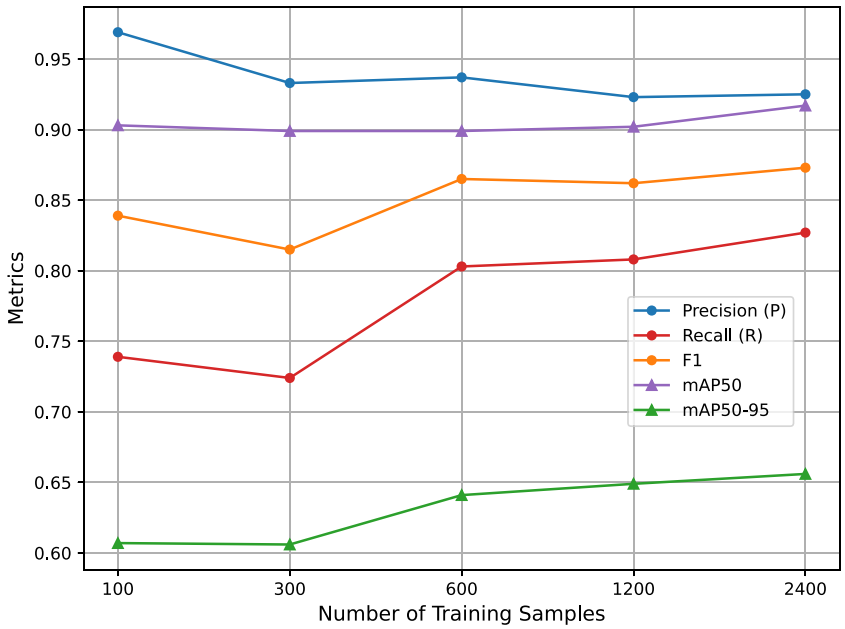

FIGURE 4 | The crop detection results of YOLOv8 trained on different numbers of training samples generated in simulation.

TABLE 3 | The crop detection results of YOLOv8 trained on simulation images of different crop species.

Crop	P	R	mAP50	mAP50-95
Sugar beet	0.909	0.802	0.889	0.612
Polygonum	0.846	0.696	0.807	0.559
Cirsium	0.966	0.616	0.885	0.585
Polygonum and Cirsium	0.196	0.746	0.725	0.52
Sugar beet and Cirsium	0.943	0.768	0.912	0.652
Sugar beet and Polygonum	0.993	0.411	0.88	0.619
All mixed	0.925	0.827	**0.917**	**0.656**

Note: The bold values indicate the best results among the compared methods for each evaluation metric.

TABLE 4 | The crop detection results of YOLOv8 trained on simulation images of sugar beet at different growth stages.

Crop	P	R	mAP50	mAP50-95
Seedling	0.886	0.552	0.748	0.452
Well-grown	0.907	0.802	0.878	0.599
Mixed	0.909	0.802	**0.889**	**0.612**

Note: The bold values indicate the best results among the compared methods for each evaluation metric.

4.3.3 | Evaluation on Background Diversity and Object Quantity

Simulation images are crucial to our framework as they offer not only annotations but also compositions for further sim-to-real transfer. To closely mimic the complexity of real-world fields, we diversified the compositions of simulation images in terms of the number of crops per image and the variety or consistency of the background across the data set. Specifically, consistent background refers to using the same soil image as background to generate all simulation images, while varied background involves randomly selecting different soil images as backgrounds for the simulations. For example, images featuring a single crop against a consistent background exhibit simpler compositions, whereas those with multiple crops against varied backgrounds display more complex compositions. Experiments using different combinations, detailed in Table 5, demonstrate that training with data of more complex compositions significantly improves the detection model's performance, underscoring that complex compositions in simulation images can greatly boost the model's generalizability to real-world scenarios.

4.4 | Cross-Evaluation on Lincolnbeet Data Set

Considering the data set's limitations in terms of clean background and influence of weeds, the Lincolnbeet data set was utilized to further demonstrate the transfer learning capabilities of the proposed training architecture Table 6. Collected with a similar configuration to our robotic system, the Lincolnbeet data set involved more challenging situations, such as varying illumination, different crop growth stages, and the presence of weeds, offering a more comprehensive assessment of the model's robustness in real-world scenarios. In the study by

TABLE 5 | The crop detection results of YOLOv8 trained on simulation images with different background (BG) diversities and object quantities.

Setting	P	R	mAP50	mAP50-95
Single Obj + consistent BG	0.512	0.466	0.516	0.339
Single Obj + varied BG	0.238	0.905	0.849	0.550
Multi Obj + consistent BG	0.607	0.927	0.910	0.613
Multi Obj + varied BG	0.925	0.827	**0.917**	**0.656**

Note: The bold values indicate the best results among the compared methods for each evaluation metric.

TABLE 6 | The mAP comparison of cross-evaluation on Lincolnbeet (LBeet) data set.

Detector	Training	Testing	mAP50
YOLOv5m (benchmark)	LBeet	LBeet	0.66
YOLOv8n (baseline)	LBeet	LBeet	0.855
YOLOv8n (ours)	DT/MARS	LBeet	0.466
YOLOv8n (ours)	DT/MARS	Adjusted LBeet	0.727

Salazar et al. (Salazar-Gomez et al. 2022), the YOLOv5m model achieved the highest mAP of 0.66, outperforming other models such as Faster R-CNN, YOLOv3, and YOLOv4 in crop detection tasks on the Lincolnbeet data set. This performance establishes the YOLOv5m model, trained with real images, as the benchmark for comparisons. We utilized this benchmark to assess the effectiveness of our detector trained with synthetic images generated by the DT/MARS-CycleGAN, aiming to gauge how well our model could simulate real-world accuracy in agricultural applications.

The YOLOv8n detector, trained on synthetic images generated by the DT/MARS-CycleGAN, was evaluated on the Lincolnbeet testing set. It achieved an mAP50 score of 0.466, which was approximately 80% of the performance of the benchmark (YOLOv5m), demonstrated strong cross-evaluation performance through in sim-to-real transfer learning, especially in different soil backgrounds and various illuminations (Figure 5). The model successfully detected most crops in the images, especially those in growth stages similar to those in the synthetic training data (Figure 5b,d–h). However, due to domain differences between the synthetic data and the Lincolnbeet data set, the detector's performance was less effective at detecting crops in the early seeding stage than in the late growth stage (Figure 5a,c). Additionally, the detector struggled to distinguish between weeds and crops, often confusing the two due to their similar shapes and appearances, which can be attributed to the lack of negative sampling of weeds in the data set (Figure 5h).

Considering the performance of crop detection could benefit from the excellent model architecture of YOLOv8, another YOLOv8n detector was trained on the Lincolnbeet data set as the baseline for further evaluation the effectiveness of synthetic images by DT/MARS-CycleGAN method. This baseline achieved an mAP50 of 0.855 on the testing set of Lincolnbeet, which was approximately double that of the detector trained solely on synthesized images. A primary factor contributing to this discrepancy is the domain gap between the Lincolnbeet data set and synthesized images, particularly in terms of the object size ratio distribution of the bounding boxes. The Lincolnbeet data set includes a significant proportion of early seedlings, accounting for about one-third of the instances categorized as small objects. However, these smaller crop types are underrepresented in the synthesized images. Addressing this issue could involve generating more synthetic images that better replicate the object size distribution found in the real data set. When objects at similar growth stages to those in the test set are considered, the synthesized image method can achieve up to 85% mean Average Precision (mAP) compared to the YOLOv8n baseline.

4.5 | Sim-to-Real Field Test—Row Detection

The sim-to-real row detection experiments were conducted on three different crops including collard, kale, and cabbage at the UGA Horticulture Research Farm in Watkinsville, GA, USA, on December 19, 2022. Each crop was planted in five rows over a 10 m × 10 m area, with a 2 m spacing between rows and 30 cm between individual plants. At the time of the experiment, the crops were approximately 5 to 10 cm in height, with most displaying between three and seven leaves (Figure 6). For image collection, we utilized the MARS-mini robot (Li et al. 2022) equipped with two RealSense D435i cameras (Intel Corporation, Santa Clara, USA.), each set at a 45° angle to provide both front and rear views. The robot, controlled remotely, traversed the fields at a consistent velocity of 0.2 m/s, and captured images at a resolution of 640 × 480 and a fps of 15 for offline analysis.

To evaluate the sim-to-real performance in the real field scenarios, we chose the crop line extraction as the downstream task after crop detection (Guo et al. 2024), which could provide the direction reference of movement in the robot's visual navigation, as well as providing supplemental visual constraints for a multi-sensor fusion navigation approach (Wang et al. 2022). The crop line in the image could be described as two offset parameters to provide the visual guidance signals depicting the error offset from a reference (Li et al. 2022; Yang et al. 2024), as illustrated in Figure 7.

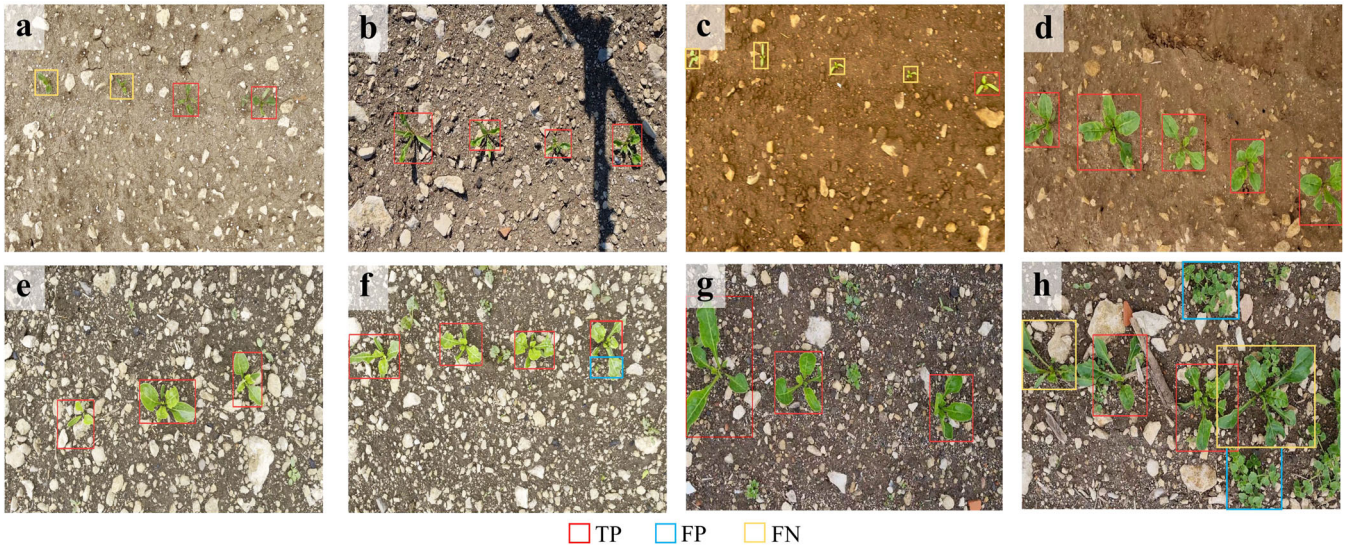

FIGURE 5 | Cross-evaluation on more diverse situations of Lincolnbeet data set, including different growth stages, soil backgrounds, illumination, weeds, and occlusion. From (a) to (h), there are four distinct lighting conditions and soil backgrounds. (a) and (c) show early growth stages, while (b) and (d) depict well-grown plants; (e) and (f) illustrate the method's accuracy against noisy backgrounds, and (g) and (h) demonstrate its effectiveness in weedy conditions with occlusions. [Color figure can be viewed at wileyonlinelibrary.com]

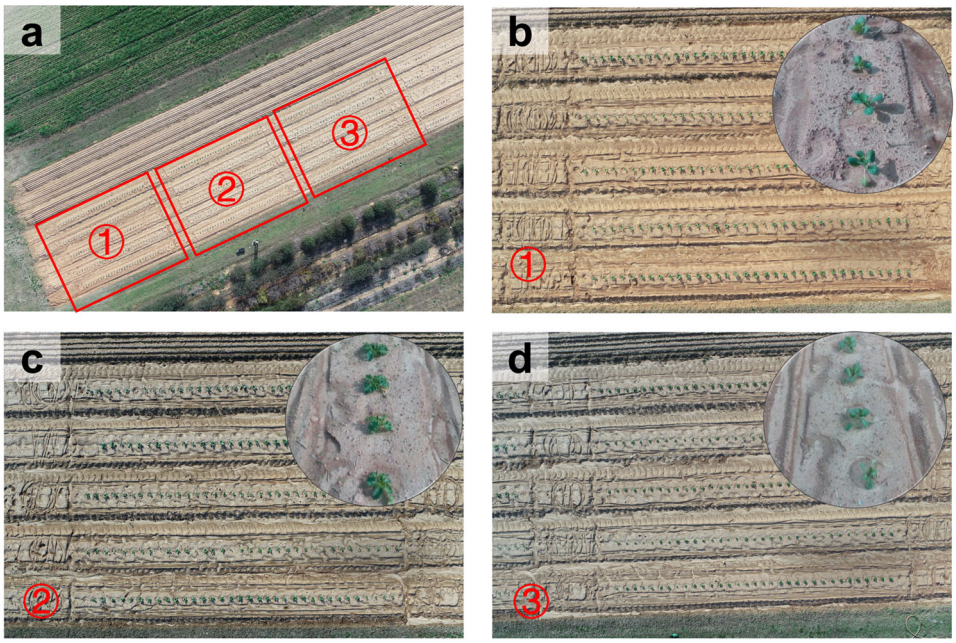

FIGURE 6 | Illustration of sim-to-real field experiment. (a) shows the overview of the experiment field; (b), (c) and (d) are the fields of collard, kale, and cabbage, respectively. [Color figure can be viewed at wileyonlinelibrary.com]

Angular_offset θ: represents the angle at which the detected crop line that deviates from the vertical orientation.

Horizontal_offset d: denotes the discrepancy between the image's central point and the detected line's center point.

In the row detection evaluation, we sampled the testing frames at a rate of 1 fps from videos collected across three crop fields. The number of sampled images for the three plots were 405, 456, and 428, respectively. For each of these images, the two offset parameters of the crop rows were manually labeled to provide a ground truth for comparison. Then, the four trained models with synthesized data (Sim-only, CycleGAN, RetinaGAN, and DT/MARS-CycleGAN) were used for crop detection and to calculate the central points of each crop. And the line fitting algorithm with the Random Sample Consensus method (RANSAC) was used to calculate the two offset parameters θ and d. Finally, we calculated the heading error and cross-track error which were the difference of angluar_-offset and horizontal_offset between the ground truth and the predictions.

Through comparing the density curve of heading error and cross-track error, the DT/MARS method demonstrated the

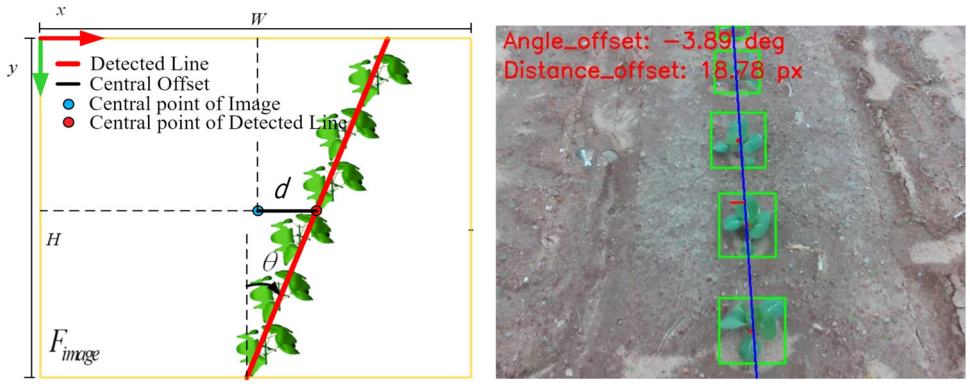

FIGURE 7 | Illustration of the row detection algorithm. Left panel: The two primary parameters used to evaluate row detection: angular offset (θ) and distance offset (d). Right Panel: An example of row detection in a real farm field. [Color figure can be viewed at wileyonlinelibrary.com]

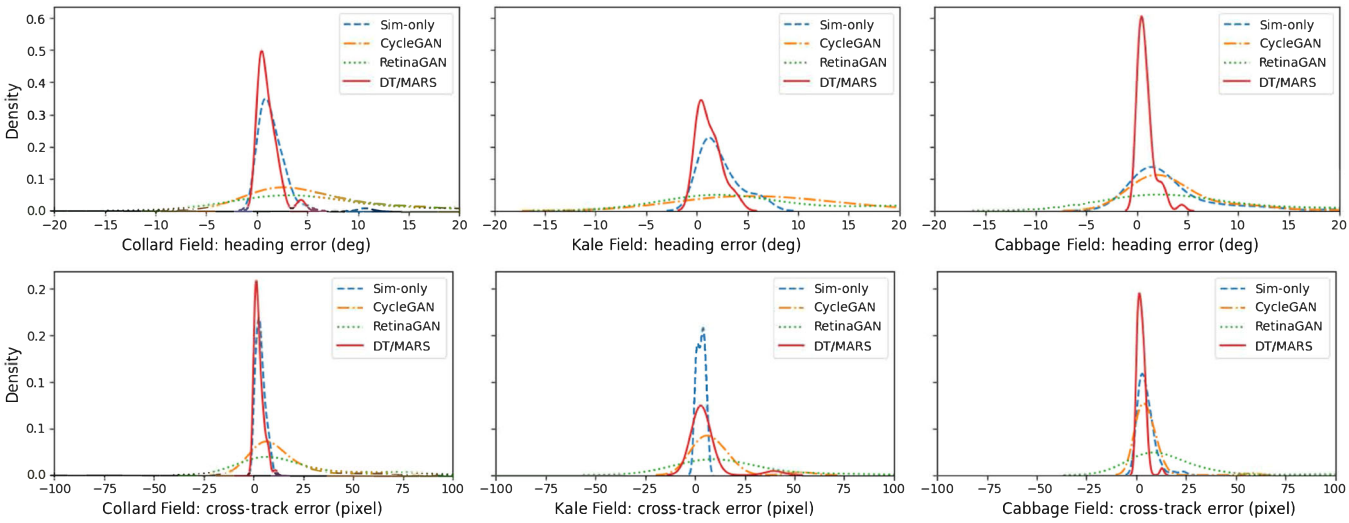

FIGURE 8 | Comparisons of DT/MARS sim-to-real with three other methods in heading errors and cross-track errors in three different vegetable crop (collard, kale, and cabbage) fields. The top row shows the heading error and the bottom row shows the cross-track error. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 7 | Mean absolute error (MAE) of heading and cross-track error across different fields.

	Collard field		Kale field		Cabbage field	
Method	**θ(°)**	**d(px)**	**θ(°)**	**d(px)**	**θ(°)**	**d(px)**
Sim-only	1.49	3.70	2.25	**2.60**	3.49	4.63
CycleGAN	7.00	17.10	9.45	11.95	4.14	6.60
RetinaGAN	10.84	32.18	7.95	28.82	8.95	27.84
DT/MARS	**1.03**	**2.66**	**1.21**	4.06	**0.85**	**2.44**

Note: The bold values indicate the best results among the compared methods for each evaluation metric.

best sim2real ability with the smallest gap to the ground truth, as depicted in Figure 8 and Table 7. The error curves of the DT/MARS method are tightly clustered around the zero mark for both parameters across all three crops, indicating a minimal deviation from the ground truth. In most of the error density curves, DT/MARS method demonstrated the highest peak that was close to the zero error with the tightest spread, as well as the most central tendency. It suggested that the detector trained with synthetic images generated by the DT/MARS method was highly effective in capturing and reproducing the actual geometrical configuration of the fields. The sim-only method, while outperforming the other two GAN-based methods (CycleGAN and Retina), even demonstrated a smaller distance difference than the DT/MARS method in the kale field, underscoring the efficacy of simulation methods in realizing sim2real transfer. In this context, the proposed DT/MARS method further outperformed Sim-only, indicating its significant potential to further bridge the gap between simulation and reality by generating more realistic imagery.

5 | Conclusions and Discussions

This paper introduces a novel DT/MARS-CycleGAN framework, effectively bridging the reality gap between simulated and real-world environments and facilitating effective zero-shot sim-to-real transfer in robotic crop detection. By imposing both cycle and DT-MARS consistency losses, which penalizes discrepancies in visual appearance and crop target between synthesized and real-world crop images, the proposed method achieves highly effective sim-to-real transfer. The fine-tuned object detectors on this diverse and high-quality synthetic data substantially elevate detection performance. Extensive experiments demonstrate that the proposed DT/MARS-CycleGAN framework enables more robust crop perception from complex backgrounds, advancing the field of robotic crop phenotyping. This work provides an effective solution to a critical robotic vision challenge in unstructured and complex agricultural environments, moving towards a more productive and resilient crop phenotyping and crop breeding process in the future.

Although our approach is promising, we acknowledge a few limitations for future improvements. The crop species, growth stage, and image background used in our data set are limited in complexity. To enhance the model's generalization capabilities, a higher degree of variability in the data set needs to be tested in the future. For example, the enriched data set should include more crop species, weeds, later growth stages, different environmental stresses, and samples from varying farming practices, capturing the full scope of variability inherent in agricultural settings. This work can be further extended along the general direction of foundation models and large vision models in the future (Bommasani et al. 2021; Zhou et al. 2023). Though the DT/MARS-CycleGAN framework demonstrated a reasonably good zero-shot transfer capability, its performance in dealing with the variation of real-world crop images can be further improved using the new methodology of foundation models (Bommasani et al. 2021). To realize such a goal, the backbones of DT/MARS-CycleGAN and object detection networks would be upgraded by the more powerful ViT model and its variants (Dosovitskiy et al. 2020; Chen et al. 2022; Kirillov et al. 2023), and should be trained and fine-tuned on much larger datasets, towards a foundation model of crop image understanding and other tasks in the future.

Data Availability Statement

The data that support the findings of this study are openly available in DT/MARS CycleGAN for Crops detection at https://www.kaggle.com/datasets/zhengkunli3969/dtmars-cyclegan.

Endnotes

[1] https://pybullet.org.

References

Ahmadi, A., M. Halstead, and C. McCool. 2021. " Towards Autonomous Crop-Agnostic Visual Navigation in Arable Fields." *arXiv preprint arXiv:2109.11936*.

Alves, R. G., G. Souza, R. F. Maia, et al. 2019. "A Digital Twin for Smart Farming." In *2019 IEEE Global Humanitarian Technology Conference (GHTC), Seattle, WA, USA*, 1–4. https://doi.org/10.1109/GHTC46095.2019.9033075.

Angin, P., M. H. Anisi, F. Göksel, C. Gürsoy, and A. Büyükgülcü. 2020. "Agrilora: A Digital Twin Framework for Smart Agriculture." *Journal of Wireless Mobile Networks Ubiquitous Computing and Dependable Applications* 11, no. 4: 77–96.

Atefi, A., Y. Ge, S. Pitla, and J. Schnable. 2021. "Robotic Technologies for High-Throughput Plant Phenotyping: Contemporary Reviews and Future Perspectives." *Frontiers in Plant Science* 12: 611940. https://doi.org/10.3389/fpls.2021.611940.

Bai, Y., B. Zhang, N. Xu, J. Zhou, J. Shi, and Z. Diao. 2023. "Vision-Based Navigation and Guidance for Agricultural Autonomous Vehicles and Robots: A Review." *Computers and Electronics in Agriculture* 205, no. 6: 107584.

Barth, R., J. Hemming, and E. J. Van Henten. 2020. "Optimising Realism of Synthetic Images Using Cycle Generative Adversarial Networks for Improved Part Segmentation." *Computers and Electronics in Agriculture* 173, no. 6: 105378.

Bi, L., and G. Hu. 2020. "Improving Image-Based Plant Disease Classification With Generative Adversarial Network under Limited Training Set." *Frontiers in Plant Science* 11: 583438. https://doi.org/10.3389/fpls.2020.583438.

Bommasani, R., D. A. Hudson, E. Adeli, et al. 2021. " On the Opportunities and Risks of Foundation Models." *arXiv preprint arXiv:2108.07258*.

Bousmalis, K., N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. 2017. "Unsupervised Pixel-Level Domain Adaptation With Generative Adversarial Networks." *CVPR* 17: 3722–3731.

Brock, A. 2018. " Large Scale Gan Training for High Fidelity Natural Image Synthesis." *arXiv preprint arXiv:1809.11096*.

Chawade, A., J. van Ham, H. Blomquist, O. Bagge, E. Alexandersson, and R. Ortiz. 2019. "High-Throughput Field-Phenotyping Tools for Plant Breeding and Precision Agriculture." *Agronomy* 9, no. 5: 258. https://doi.org/10.3390/agronomy9050258.

Chebrolu, N., P. Lottes, A. Schaefer, W. Winterhalter, W. Burgard, and C. Stachniss. 2017. "Agricultural Robot Dataset for Plant Classification, Localization and Mapping on Sugar Beet Fields." *International Journal of Robotics Research* 36, no. 10: 1045–1052.

Chen, Y., Z. Xiao, L. Zhao, et al. 2022. " Mask-Guided Vision Transformer (mg-vit) for Few-Shot Learning." *arxiv*. https://arxiv.org/abs/2205.09995.

De Silva, R., G. Cielniak, G. Wang, and J. Gao. 2024. "Deep Learning-Based Crop Row Detection for Infield Navigation of Agri-Robots." *Journal of Field Robotics* 41: 2299–2321. https://doi.org/10.1002/rob.22238.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, et al. 2020. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *arXiv preprint arXiv: 2010.11929*.

Drees, L., L. V. Junker-Frohn, J. Kierdorf, and R. Roscher. 2021. "Temporal Prediction and Evaluation of Brassica Growth in the Field Using Conditional Generative Adversarial Networks." *Computers and Electronics in Agriculture* 190: 106415.

Esser, P., R. Rombach, and B. Ommer. 2021. "Taming Transformers for High-Resolution Image Synthesis." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA*, 12868–12878. https://doi.org/10.1109/CVPR46437.2021.01268.

Fischler, M. A., and R. C. Bolles. 1981. "Random Sample Consensus: A Paradigm for Model Fitting With Applications to Image Analysis and Automated Cartography." *Communications of the ACM* 24, no. 6: 381–395.

Girshick, R., J. Donahue, T. Darrell, and J. Malik. 2014. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." In *2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA*, 580–587. https://doi.org/10.1109/CVPR.2014.81.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, et al. 2020. "Generative Adversarial Networks." *Communications of the ACM* 63: 139–144.

Guo, P., Z. Diao, C. Zhao, et al. 2024. "Navigation Line Extraction Algorithm for Corn Spraying Robot Based on YOLOv8s-Cornnet." *Journal of Field Robotics* 41: 1887–1899. https://doi.org/10.1002/rob.22360.

Hartley, Z. K., and A. P. French. 2021. "Domain Adaptation of Synthetic Images for Wheat Head Detection." *Plants* 10, no. 12: 2633.

He, L., W. Fang, G. Zhao, et al. 2022a. "Fruit Yield Prediction and Estimation in Orchards: A State-of-the-Art Comprehensive Review for Both Direct and Indirect Methods." *Computers and Electronics in Agriculture* 195: 106812.

He, L., W. Fang, G. Zhao, et al. 2022b. "Fruit Yield Prediction and Estimation in Orchards: A State-of-the-Art Comprehensive Review for Both Direct and Indirect Methods." *Computers and Electronics in Agriculture* 195: 106812.

Ho, D., K. Rao, Z. Xu, E. Jang, M. Khansari, and Y. Bai. 2021. "Retinagan: An Object-Aware Approach to Sim-to-Real Transfer." In *2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China*, 10920–10926. https://doi.org/10.1109/ICRA48506.2021.9561157.

Höfer, S., K. Bekris, A. Handa, et al. 2020. " Perspectives on Sim2real Transfer for Robotics: A Summary of the r: Ss 2020 Workshop." *arXiv preprint arXiv:2012.03806*.

Huang, P., L. Zhu, Z. Zhang, and C. Yang. 2021. "Row End Detection and Headland Turning Control for an Autonomous Banana-Picking Robot." *Machines* 9, no. 5: 103.

Huang, Z., Y. Shen, J. Li, M. Fey, and C. Brecher. 2021. "A Survey on AI-Driven Digital Twins in Industry 4.0: Smart Manufacturing and Advanced Robotics." *Sensors* 21, no. 19: 6340.

Iqbal, J., R. Xu, H. Halloran, and C. Li. 2020. "Development of a Multi-Purpose Autonomous Differential Drive Mobile Robot for Plant Phenotyping and Soil Sensing." *Electronics* 9, no. 9: 1550.

Jayaraman, P. P., A. Yavari, D. Georgakopoulos, A. Morshed, and A. Zaslavsky. 2016. "Internet of Things Platform for Smart Farming: Experiences and Lessons Learnt." *Sensors* 16, no. 11: 1884.

Jiang, Y., and C. Li. 2020. "Convolutional Neural Networks for Image-Based High-Throughput Plant Phenotyping: A Review." *Plant Phenomics* 2020: 4152816.

Jones, D., C. Snider, A. Nassehi, J. Yon, and B. Hicks. 2020. "Characterising the Digital Twin: A Systematic Literature Review." *CIRP Journal of Manufacturing Science and Technology* 29: 36–52.

Kampker, A., V. Stich, P. Jussen, B. Moser, and J. Kuntz. 2019. "Business Models for Industrial Smart Services-The Example of a Digital Twin for a Product-Service-System for Potato Harvesting." *Procedia Cirp* 83: 534–540.

Karam, C., M. Awad, Y. AbouJawdah, N. Ezzeddine, and A. Fardoun. 2022. "Gan-Based Semi-Automated Augmentation Online Tool for Agricultural Pest Detection: A Case Study on Whiteflies." *Frontiers in Plant Science* 13: 813050.

Karras, T., S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. 2020. "Analyzing and Improving the Image Quality of Stylegan." In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA*, 8107–8116. https://doi.org/10.1109/CVPR42600.2020.00813.

Katyara, S., F. Ficuciello, D. G. Caldwell, F. Chen, and B. Siciliano. 2021. "Reproducible Pruning System on Dynamic Natural Plants for Field Agricultural Robots." In *Human-Friendly Robotics 2020: Springer Proceedings in Advanced Robotics*, Vol 18, edited by M. Saveriano, E. Renaudo, A. Rodríguez-Sánchez, and J. Piater. Cham: Springer. https://doi.org/10.1007/978-3-030-71356-0_1.

Khalifa, N. E., M. Loey, and S. Mirjalili. 2022. "A Comprehensive Survey of Recent Trends in Deep Learning for Digital Images Augmentation." *Artificial Intelligence Review* 55: 2351–2377.

Kirillov, A., E. Mintun, N. Ravi, et al. 2023. " Segment Anything." *arXiv preprint arXiv:2304.02643*.

Kleeberger, K., R. Bormann, W. Kraus, and M. F. Huber. 2020. "A Survey on Learning-Based Robotic Grasping." *Current Robotics Reports* 1, no. 4: 239–249.

Li, X., W. Wang, L. Wu, et al. 2020. "Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection." *Advances in Neural Information Processing Systems* 33: 21002–21012.

Li, Z., R. Xu, C. Li, and L. Fu. 2022. "Simulation of an In-Field Phenotyping Robot: System Design, Vision-Based Navigation and Field Mapping." In *2022 ASABE Annual International Meeting*, 1. St. Joseph, MI: American Society of Agricultural and Biological Engineers.

Liang, X., B. Chen, C. Wei, and X. Zhang. 2022. "Inter-Row Navigation Line Detection for Cotton With Broken Rows." *Plant Methods* 18, no. 1: 90.

Lin, T. -Y., M. Maire, S. Belongie, et al. 2014. "Microsoft Coco: Common Objects in Context." In *Computer Vision-ECCV 2014: Lecture Notes in Computer Science* Vol 8693, edited by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Cham: Springer. https://doi.org/10.1007/978-3-319-10602-1_48.

Liu, D., Y. Chen, and Z. Wu. 2023. "Digital Twin (dt)-Cyclegan: Enabling Zero-Shot Sim-to-Real Transfer of Visual Grasping Models." *IEEE Robotics and Automation Letters* 8, no. 5: 2421–2428.

Liu, M., S. Fang, H. Dong, and C. Xu. 2021. "Review of Digital Twin About Concepts, Technologies, and Industrial Applications." *Journal of Manufacturing Systems* 58: 346–361.

Liu, Y. 2022. "A Digital Twin-Based Sim-to-Real Transfer for Deep Reinforcement Learning-Enabled Industrial Robot Grasping." *Robotics and Computer-Integrated Manufacturing* 78: 102365.

Lu, Y., D. Chen, E. Olaniyi, and Y. Huang. 2022. "Generative Adversarial Networks (GANS) for Image Augmentation in Agriculture: A Systematic Review." *Computers and Electronics in Agriculture* 200: 107208.

Lu, Y., and S. Young. 2020a. "A Survey of Public Datasets for Computer Vision Tasks in Precision Agriculture." *Computers and Electronics in Agriculture* 178: 105760.

Lu, Y., and S. Young. 2020b. "A Survey of Public Datasets for Computer Vision Tasks in Precision Agriculture." *Computers and Electronics in Agriculture* 178: 105760.

Madsen, S. L., M. Dyrmann, R. N. Jørgensen, and H. Karstoft. 2019. "Generating Artificial Images of Plant Seedlings Using Generative Adversarial Networks." *Biosystems Engineering* 187: 147–159.

Madsen, S. L., A. K. Mortensen, R. N. Jørgensen, and H. Karstoft. 2019. "Disentangling Information in Artificial Images of Plant Seedlings Using Semi-Supervised Gan." *Remote Sensing* 11, no. 22: 2671.

Moghadam, P., T. Lowe, and E. J. Edwards. 2020. "Digital Twin for the Future of Orchard Production Systems." *Proceedings* 36: 92.

Patel, V. M., R. Gopalan, R. Li, and R. Chellappa. 2015. "Visual Domain Adaptation: A Survey of Recent Advances." *IEEE Signal Processing Magazine* 32, no. 3: 53–69.

Peng, X. B., M. Andrychowicz, W. Zaremba, and P. Abbeel. 2018. "Sim-to-Real Transfer of Robotic Control With Dynamics Randomization." In *2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia*, 3803–3810. https://doi.org/10.1109/ICRA.2018.8460528.

Polvara, R., S. Molina, I. Hroob, et al. 2024. "Bacchus Long-Term (BLT) Data Set: Acquisition of the Agricultural Multimodal Blt Data Set With Automated Robot Deployment." *Journal of Field Robotics* 41: 2280.

Purcell, W., and T. Neubauer. 2023. "Digital Twins in Agriculture: A State-of-the-Art Review." *Smart Agricultural Technology* 3: 100094.

Pylianidis, C., S. Osinga, and I. N. Athanasiadis. 2021. "Introducing Digital Twins to Agriculture." *Computers and Electronics in Agriculture* 184: 105942.

Qiao, Y., J. Valente, D. Su, Z. Zhang, and D. He. 2022. "Editorial: AI, Sensors and Robotics in Plant Phenotyping and Precision Agriculture." *Frontiers in Plant Science* 13: 1064219.

Rahaman, M. M., D. Chen, Z. Gillani, C. Klukas, and M. Chen. 2015. "Advanced Phenotyping and Phenotype Data Analysis for the Study of Plant Growth and Development." *Frontiers in Plant Science* 6: 619.

Rao, K., C. Harris, A. Irpan, S. Levine, J. Ibarz, and M. Khansari. 2020. "Rl-Cyclegan: Reinforcement Learning Aware Simulation-to-Real." In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA*, 11154–11163. https://doi.org/10.1109/CVPR42600.2020.01117.

Sa, I., Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool. 2016. "Deepfruits: A Fruit Detection System Using Deep Neural Networks." *Sensors* 16: 8.

Salazar-Gomez, A., M. Darbyshire, J. Gao, E. I. Sklar, and S. Parsons. 2022. "Beyond Map: Towards Practical Object Detection for Weed Spraying in Precision Agriculture." In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan*, 9232–9238. https://doi.org/10.1109/IROS47612.2022.9982139.

Salimans, T., I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. 2016. "Improved Techniques for Training Gans." *Advances in Neural Information Processing Systems* 29: 2234.

Shete, S., S. Srinivasan, and T. A. Gonsalves. 2020. "Tasselgan: An Application of the Generative Adversarial Model for Creating Field-Based Maize Tassel Data." *Plant Phenomics* 2020: 8309605.

Solawetz, J. 2023. "What is Yolov8? The Ultimate Guide." https://blog.roboflow.com/whats-new-in-yolov8/.

Tian, H., T. Wang, Y. Liu, X. Qiao, and Y. Li. 2020. "Computer Vision Technology in Agricultural Automation—A Review." *Information Processing in Agriculture* 7, no. 1: 1–19.

Truong, J., M. Rudolph, N. H. Yokoyama, S. Chernova, D. Batra, and A. Rai. 2023. "Rethinking Sim2real: Lower Fidelity Simulation Leads to Higher Sim2real Transfer in Navigation." In *Conference on Robot Learning*, 859–870. PMLR.

Vierbergen, W., A. Willekens, D. Dekeyser, S. Cool, et al. 2023. "Sim2real Flower Detection Towards Automated Calendula Harvesting." *Biosystems Engineering* 234: 125–139.

Wang, T., B. Chen, Z. Zhang, H. Li, and M. Zhang. 2022. "Applications of Machine Vision in Agricultural Robot Navigation: A Review." *Computers and Electronics in Agriculture* 198: 107085.

Weyler, J., F. Magistri, E. Marks, et al. 2024. "Phenobench: A Large Dataset and Benchmarks for Semantic Image Interpretation in the Agricultural Domain." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46: 9583.

Winterhalter, W., F. Fleckenstein, C. Dornhege, and W. Burgard. 2021. "Localization for Precision Navigation in Agricultural Fields-Ťbeyond Crop Row Following." *Journal of Field Robotics* 38, no. 3: 429–451.

Xu, Q., G. Huang, Y. Yuan, et al. 2018. " An Empirical Study on Evaluation Metrics of Generative Adversarial Networks." *arXiv preprint arXiv:1806.07755*.

Xu, R., and C. Li. 2022a. "A Modular Agricultural Robotic System (MARS) for Precision Farming: Concept and Implementation." *Journal of Field Robotics* 39, no. 4: 387–409.

Xu, R., and C. Li. 2022b. "A Review of High-Throughput Field Phenotyping Systems: Focusing on Ground Robots." *Plant Phenomics* 2022: 9760269.

Yang, M., C. Huang, Z. Li, et al. 2024. "Autonomous Navigation Method Based on RGB-D Camera for a Crop Phenotyping Robot." *Journal of Field Robotics* 41: 2663.

Zhang, S., Y. Liu, K. Xiong, et al. 2024. "A Review of Vision-Based Crop Row Detection Method: Focusing on Field Ground Autonomous Navigation Operations." *Computers and Electronics in Agriculture* 222: 109086.

Zheng, Z., P. Wang, W. Liu, J. Li, R. Ye, and D. Ren. 2020. "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression." *Proceedings of the AAAI Conference on Artificial Intelligence* 34, no. 7: 12993–13000.

Zhou, C., Q. Li, C. Li, et al. 2023. "A Comprehensive Survey on Pretrained Foundation Models: A History From BERT to ChatGpt." https://arxiv.org/abs/2108.07258.

Zhu, F., M. He, and Z. Zheng. 2020. "Data Augmentation Using Improved cDCGAN for Plant Vigor Rating." *Computers and Electronics in Agriculture* 175: 105603.

Zhu, J.-Y., T. Park, P. Isola, and A. A. Efros. 2017. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks." In *2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017*, 2242–2251. https://doi.org/10.1109/ICCV.2017.244.