Journal of *Imaging*

MDPI

Article

Efficient Wheat Head Segmentation with Minimal Annotation: A Generative Approach

Jaden Myers [1], Keyhan Najafian [2], Farhad Maleki [1] and Katie Ovens [1,*]

[1] Department of Computer Science, University of Calgary, 2500 University Drive NW, Calgary, AB T2N 1N4, Canada; jaden.myers@ucalgary.ca (J.M.); farhad.maleki1@ucalgary.ca (F.M.)
[2] Department of Computer Science, University of Saskatchewan, Saskatoon, SK S7N 5A2, Canada; keyhan.najafian@usask.ca
* Correspondence: katie.ovens@ucalgary.ca

Abstract: Deep learning models have been used for a variety of image processing tasks. However, most of these models are developed through supervised learning approaches, which rely heavily on the availability of large-scale annotated datasets. Developing such datasets is tedious and expensive. In the absence of an annotated dataset, synthetic data can be used for model development; however, due to the substantial differences between simulated and real data, a phenomenon referred to as domain gap, the resulting models often underperform when applied to real data. In this research, we aim to address this challenge by first computationally simulating a large-scale annotated dataset and then using a generative adversarial network (GAN) to fill the gap between simulated and real images. This approach results in a synthetic dataset that can be effectively utilized to train a deep-learning model. Using this approach, we developed a realistic annotated synthetic dataset for wheat head segmentation. This dataset was then used to develop a deep-learning model for semantic segmentation. The resulting model achieved a Dice score of 83.4% on an internal dataset and Dice scores of 79.6% and 83.6% on two external datasets from the Global Wheat Head Detection datasets. While we proposed this approach in the context of wheat head segmentation, it can be generalized to other crop types or, more broadly, to images with dense, repeated patterns such as those found in cellular imagery.

Keywords: deep learning; segmentation; generative adversarial networks; data synthesis

check for updates

Citation: Myers, J.; Najafian, K.; Maleki, F.; Ovens, K. Efficient Wheat Head Segmentation with Minimal Annotation: A Generative Approach. *J. Imaging* **2024**, *10*, 152. https://doi.org/10.3390/jimaging10070152

Academic Editors: Pierre Gouton, Saeid Minaei and Vahid Mohammadi

Received: 10 April 2024
Revised: 4 June 2024
Accepted: 11 June 2024
Published: 21 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Deep learning models have shown great potential for semantic image segmentation [1,2], where a label is assigned to each pixel representing its semantics. Most semantic segmentation models are developed using supervised approaches that rely on the availability of large amounts of annotated data. However, gathering large amounts of annotated data needed to train these models is often expensive, time-consuming [3], and potentially requires highly specialized expertise to provide the image annotation. The need for a large annotated data set can be alleviated by using synthetic annotated data. However, a model trained on simple synthetic data is unlikely to perform well when applied to real-world data. Therefore, effective integration of synthetic data into segmentation tasks requires overcoming the inherent distribution shift between synthesized and real images, thereby demanding domain adaptation strategies to bridge this gap.

Segmentation has been applied to annotate different plant components to detect abnormalities in crops such as lodging [4], and important phenotypic plant traits in precision agriculture, such as organ size [5], organ health [6], and response to biotic [7] and abiotic stress [8]. Large annotated datasets, such as the Global Wheat Head Detection (GWHD) dataset [9], have enabled the development of new supervised deep learning-based methods for plant phenotyping from field images. Since the pixel-level annotation required for

segmentation tasks is substantially more demanding due to the time and effort invested, most of these datasets are focused on classification or object detection tasks and rarely on segmentation tasks.

Some recently proposed solutions for wheat head segmentation aim to accomplish this task with little manual annotation [10–12]. In these cases, synthesizing new samples to improve model performance for segmentation tasks is essential, but has the added challenge of distribution shift between the synthesized and real images and requires domain adaptation steps to gradually bridge the domain gap.

Fourati et al. utilized Faster R-CNN and EfficientDet models trained on the GWHD dataset to develop a wheat head detection method [10]. They used semi-supervised techniques such as pseudo-labeling, test time augmentation, multi-scale ensemble, and bootstrap aggregation to improve model performance, and achieved a mean average precision of 0.74. Najafian et al. generated a synthesized dataset using a cut-and-paste approach and trained a YOLO architecture for wheat head detection [11]. Fine-tuning the model with the GWHD dataset resulted in a mean average precision of 0.82. However, their model showed lower performance when trained solely on the synthesized dataset (Precision: 0.318, Recall: 0.130).

Najafian et al. proposed a semi-self-supervised approach for semantic segmentation, utilizing both computationally and manually annotated images. They synthesized an image dataset for semantic segmentation models using a wheat field video clip, a few annotated images, and background scenes [12]. They trained a customized U-Net model with this synthesized dataset and applied domain adaptation steps to address the domain gap between synthesized and real images, and achieved a Dice score of 0.89 and 0.73 on their internal and external test sets, respectively. In this approach, the distinction between the wheat heads and the background in the synthesized images is highly contrasted, making it relatively trivial for a model to learn as it can rely on changes in contrast to perform the segmentation task. Consequently, this method, without further domain adaptation techniques, is not generalizable to wheat head segmentation in real images.

We circumvent these pitfalls by training a model for the segmentation of wheat heads with a modified generative adversarial network (GAN) approach to generate the synthetic data set. A GAN traditionally consists of two subnetworks: a generative network and a discriminative network [13]. The role of the generative network is to generate realistic samples, while the discriminator model takes the generative network's output as input along with real-world data and provides a binary classification output; true if the input is real or false if it is generated. The parameters of the generative model are adjusted according to the classifications made by the discriminative model and a loss function. The parameters of the discriminative model are adjusted in a similar way based on its classification of the data. The generative and discriminative models will be pitted against each other in a zero-sum game until the discriminative model is no longer able to consistently tell the difference between the synthetic and real data. This process results in a generative network that is able to generate images that closely resemble real-world images.

Cycle-Consistent Generative Adversarial Networks (CycleGANs) are a popular GAN-based deep learning framework focused on the task of unpaired image-to-image translation [14]. For example, they can be used to transform images of a crop field captured under different lighting conditions or with different camera types. This ability to translate images from a source domain to a target domain without the need for paired images allows for data augmentation, domain adaptation, and image enhancement in agriculture applications. The fundamental idea behind CycleGAN is to learn mappings between two different domains without the need for paired training data. Traditional methods for image translation, such as Pix2Pix [15], require paired examples of corresponding images in the source and target domains. However, acquiring such paired data can be challenging and labor-intensive or impractical in many cases. CycleGAN addresses this limitation by using unpaired data for training, making it more flexible and widely applicable. In agriculture,

CycleGANs have been utilized for crop disease detection [16], plant phenotyping [17], and crop yield prediction [18].

The limitation of simply using CycleGAN for the purpose of developing segmentation models with accurate annotation is that having an exact match between the generated image and the segmentation itself will not be enforced. A potential consequence of this is that some wheat heads generated do not have a corresponding segmentation, leading to an inaccurate data annotation that will negatively affect downstream models trained on these images. Therefore, we propose to adapt CycleGAN by adding this enforcement step to maintain consistency between generated images and their corresponding annotation.

The paper is structured as follows: In Section 2, we provide an overview of the modified CycleGAN architecture employed to generate synthetic training samples for wheat head segmentation and the process for evaluating the segmentation models. Section 3 presents the results and evaluation of our proposed model. Finally, Section 4 discusses the findings, and Section 5 provides the concluding remarks.

2. Materials and Methods

This section describes the datasets, the modified CycleGAN for generating synthetic data, and the procedures for training and evaluating the segmentation models (see Figure 1).

2.1. Datasets

The data used in this study can be accessed at https://www.cs.usask.ca/ftp/pub/whs/ (accessed on 1 June 2023). All the data used to train and validate our models came from a video clip of a wheat field, hereafter referred to as W, and 11 videos of background scenes captured using Samsung cameras with resolutions of 12 and 48 Megapixels. With only a single frame from W that was manually annotated and frames from the background videos, we generated a synthetic dataset Δ_S using the cut-and-paste method proposed by Najafian et al. [12]. Figure 2 illustrates this approach. The dataset Δ_S consists of 11,000 synthetic images and their corresponding semantic segmentation masks. The rest of the frames in W, aside from 100 frames used for testing (Δ_W), formed the dataset Δ_R, consisting of real wheat field images with no semantic segmentation ground truth labels. Our goal is to use the datasets Δ_S and Δ_R to generate a dataset Δ_{GAN} of GAN-based synthetic images that more closely resemble the real images in Δ_R. Δ_{GAN} consists of 11,000 synthetic images and their corresponding semantic segmentation labels, 10,000 of which were used for training and 1000 for validation of a semantic segmentation model.

To evaluate the performance of the segmentation model trained on our dataset Δ_{GAN}, we tested its performance on 3 different evaluation data sets: (1) an internal dataset Δ_W consisting of the 100 manually annotated images from W, (2) an external dataset Δ_{GWHD} consisting of 365 samples across 18 different domains from the GWHD dataset [9] which was annotated by Najafian et al. [12] to evaluate their model, and (3) another external dataset Δ_{Tokyo} consisting of samples from the UTokyo subset of the GWHD dataset [9] which was also annotated by Najafian et al. [12] to evaluate their model. These datasets allowed us to compare the performance of our model with the models developed by Najafian et al. [12] and Rawat et al. [19]. Since the internal test set consisting of images from the same video from which we generated the dataset Δ_{GAN}, the performance of the model on the internal test set Δ_W might not be a good indication of the model's ability to generalize to new domains. However, datasets Δ_{GWHD} and Δ_{Tokyo}, as external datasets, provide a reliable evaluation of model performance and generalizability. Therefore, we used the two external test datasets—consisting of different varieties of wheat from different fields and at various stages of growth. Samples from these datasets were not seen by the model during the training process.

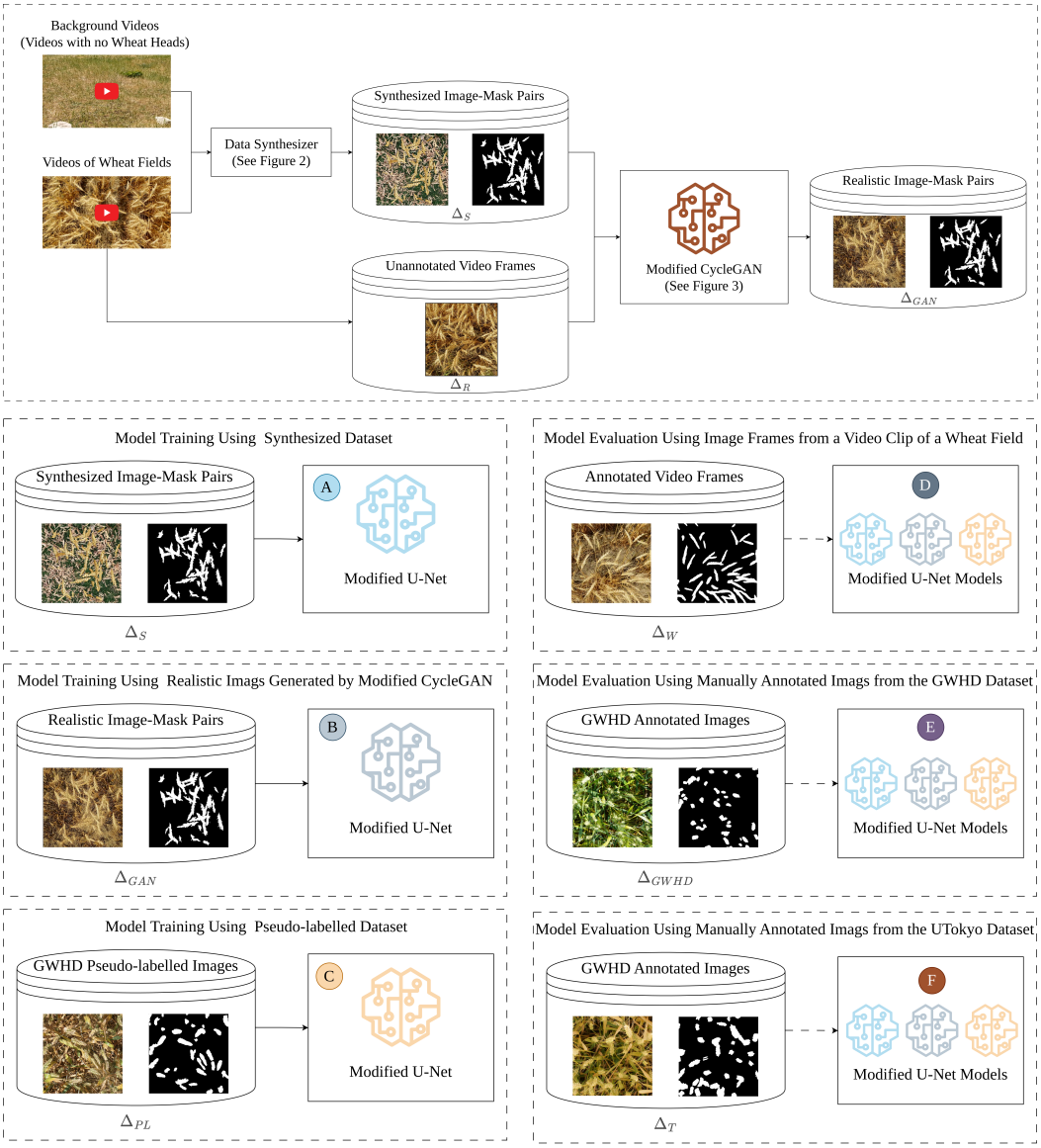


Figure 1. Schematic overview of the model development and evaluation process. The diagram in the top box illustrates a schematic overview of the proposed methodology for data synthesis using the modified CycleGAN model. Annotated images generated from the data synthesizer (see Figure 2) and real unannotated images extracted from video frames are used for development the modified CycleGAN model (see Figure 3). Boxes A, B, and C show the model training process using synthesized computationally annotated images (Δ_S), GAN-generated dataset (Δ_{GAN}), and pseudo-labeled dataset (Δ_{PL}), respectively. Boxes D, E, and F show the evaluation process for the three trained models in steps A, B, and C on the internal test set extracted from the video frames (Δ_W), the external test set from the GWHD dataset (Δ_{GWHD}), and the external test set from the UTokyo dataset (Δ_{Tokyo}).

2.2. Model Architecture

Cycle-Consistent Generative Adversarial Network (CycleGAN) architecture was designed for unpaired image-to-image translation [14]. CycleGAN consists of two generator networks and two discriminator networks. One generator transforms images from one domain to another, and the other generator does the reverse. Each generator is paired with a discriminator to distinguish real and fake images in its respective domain. In addition to an adversarial loss function, the CycleGAN model uses a cycle-consistency loss component to minimize the difference between an image and the outcome of translating the same image to another domain and then back to the original domain. By minimizing the combined loss

functions, the generator and discriminator networks can learn to perform unsupervised image translation without the need for paired training data.

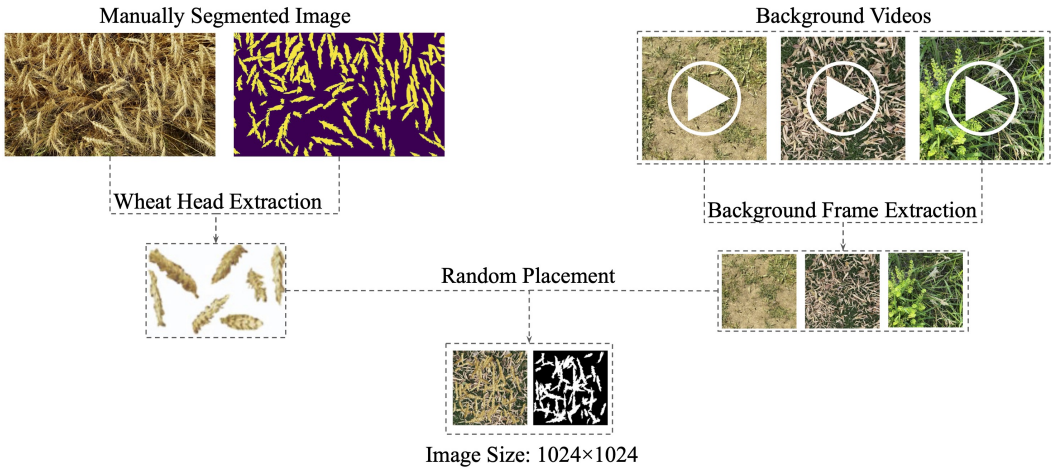


Figure 2. A visualization of the pipeline used to generate synthetic images. Wheat head cutouts are extracted from a manually annotated real wheat image and background frames are extracted from the background videos. The wheat heads are then randomly overlaid onto background frames to generate a wheat head image and its segmentation mask.

Although the unmodified CycleGAN focuses on the overall visual appearance, it does not impose any constraints to enforce semantic consistency for the masked regions during the translation process. In this paper, we utilize the copy-paste approach to computationally synthesize images alongside their segmentation mask, and then we develop a CycleGAN-inspired model architecture to transform these synthesized images into realistic images while preserving the annotations. This architecture is presented in Figure 3.

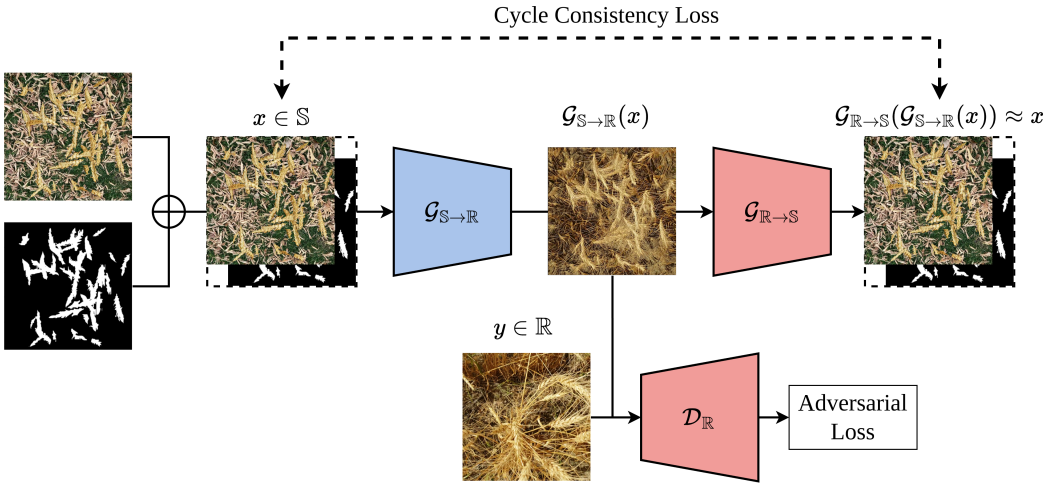


Figure 3. Diagram of the modified CycleGAN. The generator $\mathcal{G}_{\mathbb{S}\to\mathbb{R}} : \mathbb{S} \to \mathbb{R}$ takes as input synthetic images concatenated with their semantic segmentation masks $x \in S$ and outputs a corresponding real image $\mathcal{G}_{\mathbb{S}\to\mathbb{R}}(x)$. A cycle consistency loss is calculated between x and $\mathcal{G}_{\mathbb{R}\to\mathbb{S}}(\mathcal{G}_{\mathbb{S}\to\mathbb{R}}(x))$. Not present in the diagram, a cycle consistency loss is also calculated in the opposite direction with real images $y \in \mathbb{R}$ and $\mathcal{G}_{\mathbb{S}\to\mathbb{R}}(\mathcal{G}_{\mathbb{R}\to\mathbb{S}}(y))$, and the discriminator $\mathcal{D}_{\mathbb{S}}$ calculates an adversarial loss with $\mathcal{G}_{\mathbb{R}\to\mathbb{S}}(y) \in \mathbb{S}$ and x.

Given two domains $\mathbb{S}$, which represents computationally synthesized images and their segmentation masks, and $\mathbb{R}$, which represents realistic unannotated images, our aim is to transform the synthesized images to make them more realistic, while preserving the segmentation masks, such that the transformed images retain the same segmentation masks as their synthesized counterparts. We utilize a synthesized annotated dataset $\{(I_s, M_s) \mid I_s \sim \mathcal{P}_{\mathbb{S}}(I_s) \mid s = 1, \cdots, n\}$—generated based on the methodology proposed by Najafian et al. [12]—from $\mathbb{S}$ and a real unannotated dataset $\{I_r \mid I_r \sim \mathcal{P}_{\mathbb{R}}(I_r)\, r = 1, \cdots, m\}$—extracted from video frames of wheat fields—from $\mathbb{R}$. $\mathcal{P}_{\mathbb{S}}$ and $\mathcal{P}_{\mathbb{R}}$ represent the probability distribution of data from domains $\mathbb{S}$ and $\mathbb{R}$, respectively.

The proposed model architecture consists of two generator networks $\mathcal{G}_{\mathbb{S}\to\mathbb{R}}$ and $\mathcal{G}_{\mathbb{R}\to\mathbb{S}}$ and two discriminator networks $\mathcal{D}_{\mathbb{S}}$ and $\mathcal{D}_{\mathbb{R}}$. $\mathcal{G}_{\mathbb{S}\to\mathbb{R}}$ transforms a synthesized image I_s from domain $\mathbb{S}$ of synthesized images to image I_r from domain $\mathbb{R}$ of realistic images; the generator $\mathcal{G}_{\mathbb{R}\to\mathbb{S}}$ transforms I_r back to an image I'_s, which would ideally be the same as I_s. $\mathcal{D}_{\mathbb{S}}$ is paired with $\mathcal{G}_{\mathbb{R}\to\mathbb{S}}$, and $\mathcal{D}_{\mathbb{R}}$ is paired with $\mathcal{G}_{\mathbb{S}\to\mathbb{R}}$ to distinguish real and fake images in their corresponding domains. The CycleGAN model uses a cycle-consistency loss to minimize the difference between I_r and I'_r. $\mathcal{G}_{\mathbb{S}\to\mathbb{R}}$ is tasked with creating a mapping from domain $\mathbb{S}$ to domain $\mathbb{R}$ and $\mathcal{G}_{\mathbb{R}\to\mathbb{S}}$ is tasked with creating a mapping from $\mathbb{R}$ to $\mathbb{S}$.

$$\mathcal{G}_{\mathbb{S}\to\mathbb{R}} : \mathbb{S} \to \mathbb{R}$$
$$\mathcal{G}_{\mathbb{R}\to\mathbb{S}} : \mathbb{R} \to \mathbb{S}$$

First, given an input image and its segmentation mask, generator $\mathcal{G}_{\mathbb{S}\to\mathbb{R}}$ produces an image in the style of the target domain $\mathbb{R}$ without a segmentation mask. Next, generator $\mathcal{G}_{\mathbb{R}\to\mathbb{S}}$ takes the generated images and tries to regenerate the synthesized input image and its mask from domain $\mathbb{S}$. The cycle consistency loss is then calculated by measuring the distance between the original image and the recreated image, as well as the original and recreated mask. This encourages the model to preserve segmentation masks during domain translations because the masks must be accurately recreated at the end of the cycle.

After generating a realistic and computationally annotated image dataset, we develop an image segmentation model. For the segmentation model, we use the same modified U-Net [20] utilized by Najafian et al. [12]. A binary cross-entropy loss function is used to train all segmentation models. For evaluation of segmentation performance, Dice and IoU are used.

2.3. Pseudo Labelling

To enhance the performance of the segmentation model trained on our GAN-generated dataset, we introduce a domain adaptation step based on the pseudo-labeling approach [21]. From the GWHD dataset, we randomly selected 360 images that are not part of the GWHD external evaluation datasets and passed them as input to the model trained on Δ_{GAN} to obtain the predicted segmentation masks. From these predictions, we selected only the best predictions, which were then used to fine-tune the model. While highly accurate predictions were chosen manually, the effort needed to choose these predictions by verifying the quality of the predicted segmentation masks is negligible compared to the manual segmentation process. Figure 4 presents examples of images that were selected and not selected. A prediction is considered to be of high quality and should be chosen if it maximizes the area of the wheat head in the image being segmented and minimizes the non-wheat head parts of the image being segmented. A dataset of 99 pseudo-labeled images was compiled using the process.

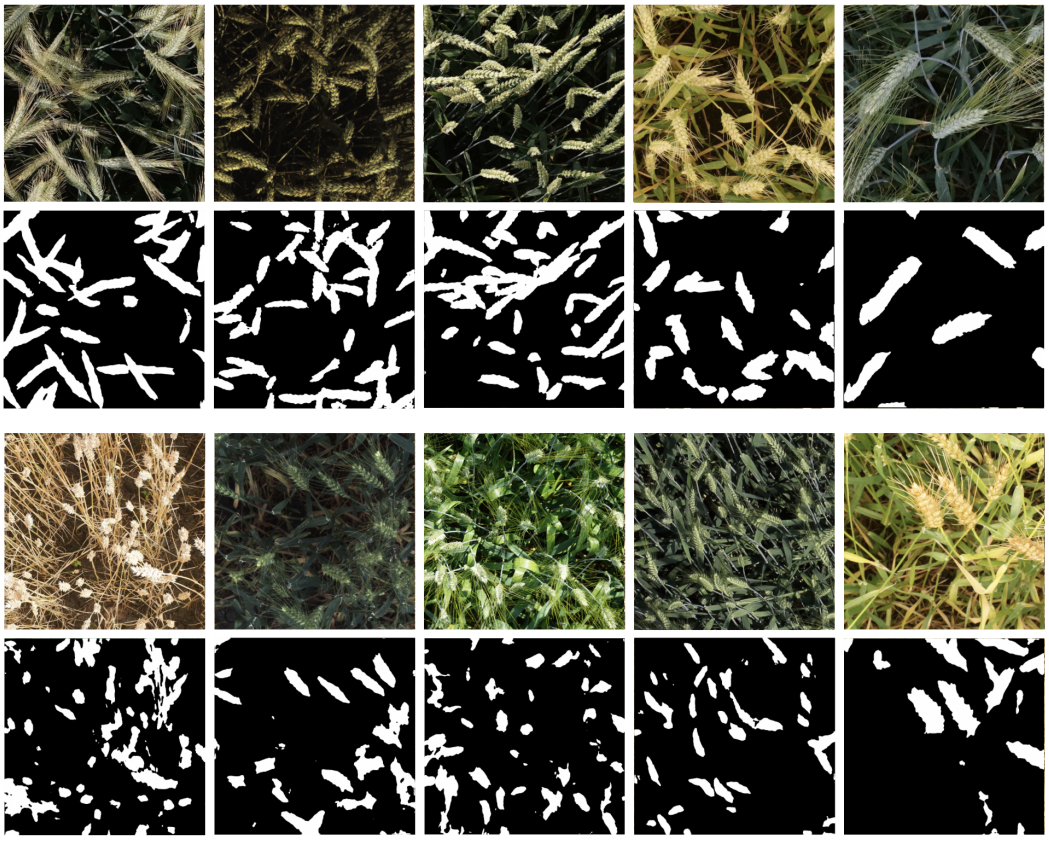


Figure 4. Examples of randomly selected GWHD images and the corresponding pseudo mask predictions. The top row consists of samples that were selected to be part of the dataset used to fine-tune the model. The bottom row consists of samples that were not selected.

2.4. Experiments

Our model development involves multiple phases which are detailed as follows. We trained our modified CycleGAN model for 60 epochs using the Adam optimizer [22] with a learning rate of 0.0002. We exploited the same optimizer for both pairs of generators and discriminators. We also utilized L1 distance as the cycle loss along with adversarial losses. We used the trained modified CycleGAN model to generate a large-scale computationally annotated dataset with the same size of Δ_S, referred to as Δ_{GAN}.

To develop the segmentation model, we closely followed the training strategy utilized by Najafian et al. [12]. We used the same model architecture and trained the model for 45 epochs on the GAN-generated data Δ_{GAN}. In the next phase, using Δ_{PL} alongside Δ_{GAN} data, we further fine-tuned the model for 45 more epochs to obtain our final model.

In all three phases of developing our segmentation model, we opted for the SGD optimizer [23] with a learning rate of 0.01, weight decay of 0.001, and momentum of 0.95. We also used Step Scheduler with a reduction rate of 0.1 in every 5 epochs. Moreover, in every step of developing our segmentation model, we evaluated the performance of the model on the validation set using the Dice score as our main score to choose the best-performing model. In each development phase, the best-performing model was evaluated on our test sets. Each model's performance is detailed and discussed in Sections 3 and 4.

3. Results

3.1. Synthesization of Wheat Head Images

Figure 5 shows randomly selected synthetic images from the dataset Δ_S, and their corresponding translated output from the modified CycleGAN in the dataset Δ_{GAN} with real wheat images for comparison. The translated synthetic images in Δ_{GAN} are visibly closer to the realistic images compared to the synthetic images from the dataset Δ_S. Moreover,

although the translated images from the original CycleGAN exhibit equivalent realism to those of our modified CycleGAN, the original CycleGAN does not preserve semantic features during the translation, as illustrated by Figure 6. Our modified CycleGAN with segmentation mask input learned to preserve the semantic features through translation as shown by Figure 7, which illustrates the model's ability to place wheat heads corresponding to the input segmentation mask.

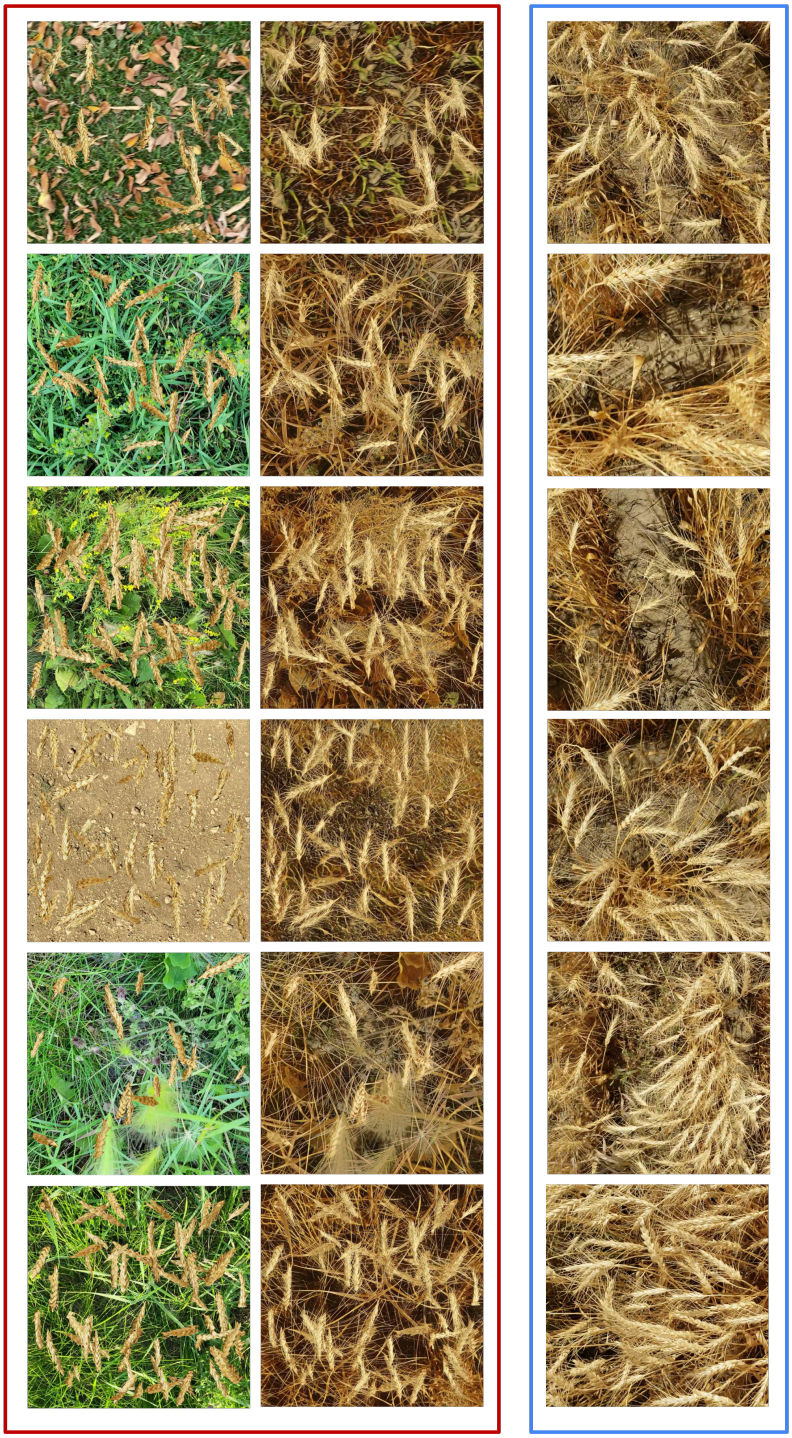


Figure 5. Examples of randomly selected synthetic images from Δ_S (left column) and their corresponding outputs from our modified CycleGAN (right column) are illustrated in the red box. The images in the blue box are randomly selected real wheat field images provided for comparison.



Figure 6. A synthetic wheat image and its corresponding output from an unmodified CycleGAN. The red circles highlight the flaws of the unmodified CycleGAN image translation, where the original CycleGAN does not preserve the same segmentation mask as the input image.

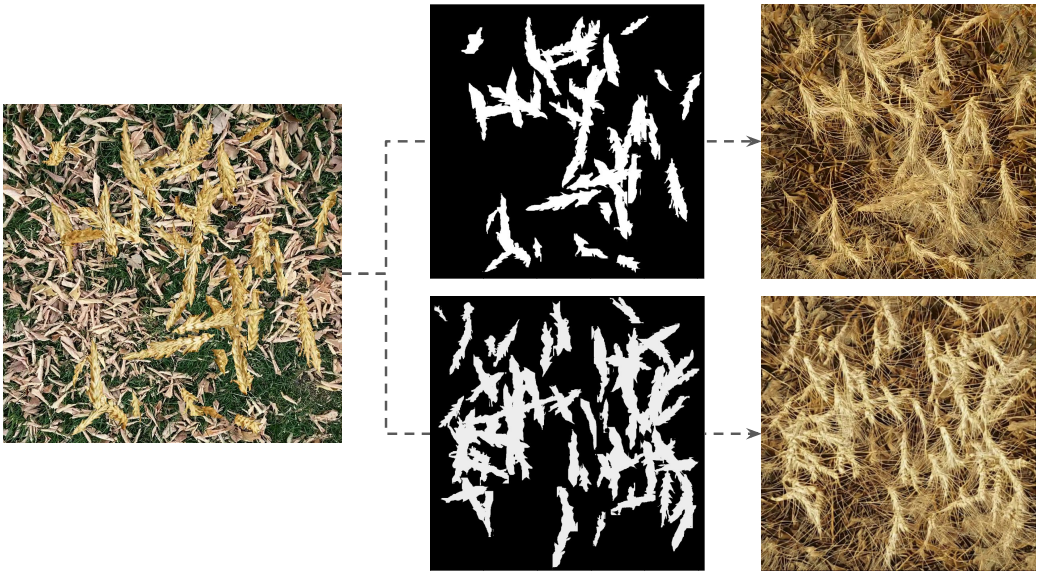


Figure 7. The same synthetic wheat image concatenated with two different semantic segmentation masks and the corresponding output from our modified CycleGAN.

3.2. Evaluation of Wheat Head Segmentation Model Trained with Generated Wheat Head Images

The performance of the segmentation models on both internal and external test sets is outlined in Table 1. Compared to the model trained on the dataset Δ_S, the model trained on our modified CycleGAN synthetic dataset Δ_{GAN} showed a significant increase of 10.2% and 12.0% in the Dice and IoU, respectively, when evaluated on the internal evaluation dataset (Δ_W). When evaluated on the external GWHD test set(Δ_{GWHD}) consisting of samples from 18 domains, the performance of the model trained on our dataset showed an even more substantial increase of 21.0% and 16.5% in the Dice and IoU scores, respectively. When evaluated on the UTokyo dataset (Δ_{Tokyo}), the model trained on the data generated from the modified CycleGAN saw an increase of 23.4% and 23.6% in Dice and IoU, respectively. Furthermore, after the model trained on the dataset Δ_{GAN} was fine-tuned using the pseudo-labeled dataset (Δ_{PL}), the model achieved an even greater increase in performance across all evaluation datasets as shown in Table 1.

Table 1. The performance of the models trained on the synthetic data sets. Model A was trained on dataset Δ_S, which is the synthesized computationally annotated dataset. Model B was trained on dataset Δ_{GAN}, which is generated by our modified CycleGAN model. Model B was fine-tuned using the pseudo-labeled dataset Δ_{PL} to generate model C.

Model	Trained On	Tested on Δ_W		Tested on Δ_{GWHD}		Tested on Δ_{Tokyo}	
		Dice	IoU	Dice	IoU	Dice	IoU
A	Δ_S	0.709	0.566	0.368	0.274	0.407	0.275
B	Δ_{GAN}	0.811	0.686	0.578	0.440	0.644	0.511
C	Δ_{GAN} + Δ_{PL}	0.834	0.720	0.796	0.687	0.836	0.754

The performance ratios between our final model and the baseline model are 1.272, 2.507, and 2.742 for the internal test set (Δ_W), GWHD dataset (Δ_{GWHD}), and UTokyo dataset (Δ_{Tokyo}), respectively. These ratios represent the IoU of model C divided by the IoU of model A.

4. Discussion

Human-annotated data acquisition is an expensive and labor-intensive task in many domains such as precision agriculture. In this work, we proposed a methodology for computationally generating large-scale labeled training data, for deep learning model development. The proposed approach allowed us to generate a synthetic training dataset, resembling real wheat field images, with little manual annotation. The model trained using this approach resulted in a much smaller domain gap to bridge when applied to real-world data. Our approach demonstrated the potential to be applied to domains, where the lack of annotated data limits the development of deep learning models.

We trained a customized U-Net model [20], utilized by Najafian et al. [12] for the segmentation task in three phases on synthetic, GAN-generated, and a small pseudo-labeled dataset. Comparing the model performance with the state-of-the-art works, our final model obtained a Dice score of 0.796 on Δ_{GWHD} test set, one of our external sets, while the best model developed by Najafian et al. [12]—trained and fine-tuned on synthetic and pseudo-labeled data—achieved a Dice score of 0.7416 using their proposed Test Time Augmentation (TTA) approach on dataset (Δ_{GWHD}). On the second external test set Δ_{Tokyo}, the best model by Najafian et al. (Model P), which they trained sequentially on the synthetic and a large-scale pseudo-labeled dataset, obtained a Dice score of 0.842. Fine-tuned on a small pseudo-labeled training set of 99 samples, our model obtained a comparable Dice score of 0.836. Moreover, our model surpassed the Rawat et al. [19] model, which was trained in a supervised manner, on the Δ_{Tokyo}. Our approach resulted in an almost 5% increase in IoU score.

4.1. Synthetic Image Generation

In the previous state-of-the-art, Najafian et al. [12] used the cut-and-paste method to synthesize computationally annotated images by overlaying extracted wheat heads on the background frames originating from videos with no wheat. Although this approach allows for the generation of large amounts of computationally annotated training data, the resulting data do not closely resemble real data. Consequently, models trained only on such data result in low performance when evaluated on real data. This phenomenon is known as domain/distribution shift. The proposed approach in this research alleviates this problem by adding a new step to the pipeline for image synthesis. We conducted image-to-image translation using a modified CycleGAN to more closely resemble real wheat field images, compared to the simple cut and paste synthetic data. The resulting generated images more closely resemble realistic wheat field images. The model trained on the resulting data, substantially improved performance when compared to initial synthetic data (see Figure 5).

When using a standard/unmodified CycleGAN model, the translated images do not preserve the semantic information, e.g., the location and shape of wheat heads. Consequently, the initial segmentation masks do not match the resulting images. Our modified CycleGAN with semantic mask input was able to generate images that closely resemble real wheat field images while preserving the semantic information of the images. Therefore, the resulting images do not need further manual annotation and the initial segmentation masks can be used along with the resulting image.

4.2. Limitations and Future Directions

To reach higher performance, our developed model can be used to predict segmentation masks for real images and the resulting masks often need little-to-no manual curation. This semi-automated approach can expedite the speed in which we develop deep learning models for wheat head segmentation and, in general, crop monitoring, as our approach is generalizable to many other crops. However, minimal manual curation is still needed for such an approach preventing us from fully automating the process. On the other hand, the direct use of pesudo-labelled data, i.e., predictions made by the model, for the further finetuning of the model can introduce noise and systematic bias into the system and is hard to reproduce, which is computationally expensive to fix when the model is trained on non-curated pseudo-labelled images. One potential avenue for future research would be to explore more automated techniques to curate pseudo-labeled data. Techniques such as uncertainty estimation could identify and prioritize the images with the least label uncertainty. Such data paired with computational approaches for image augmentations, to increase data diversity, could allow us to improve model performance without further manual curation.

In our proposed modified CycleGAN, we incorporated the recreation of segmentation masks for the purpose of preserving semantic information during image translation. This is conducted by generating an output which is the concatenation of image and mask. Utilizing a generator component for creating images and another generator model for creating the masks should be explored and the effect of such model architecture on the quality of the resulting images and masks should be studied. Also, a generator architecture such as the U-Net generator proposed by Torbunov et al. [24] could be customized for improving the image generation process. However, we expect a trade-off between the quality of the generated images and semantic preservation.

To train the modified CycleGAN for synthetic-to-real image translation, we used real images extracted from a video clip of a single wheat field. Since all real training images were extracted from the same video clip, the variation within the real training data is low. Utilizing multiple videos taken from multiple wheat fields would increase the variation within the real data and could potentially lead to the data generated by the CycleGAN being more diverse as training data, better representing the variability of real-world data. Incorporating images that bring more variety into the training data, e.g., regarding environmental conditions and growth stage of the wheat fields, would be beneficial for generating realistic images under a variety of different conditions. Having crop images at various stages of development during different growing seasons, weather conditions, and lighting conditions could help to further improve the generalizability of wheat head segmentation models.

In the initial copy–paste image synthesis, wheat heads are randomly overlaid on the background frames. This falls short in precisely emulating the layout of wheat heads in real wheat fields, and our CycleGAN approach tries to preserve this layout. The layout of the initial synthesized wheat field images could be adjusted to more closely resemble that of real images. Creating a pipeline that uses density maps to more closely resemble the rows and orientations of wheat heads in real wheat images, while accounting for variations resulting from wind and movements of crops instead of complete random placement, could further decrease the domain gap between the synthetic and real wheat data.

5. Conclusions

In this study, we proposed an approach for generating computationally annotated data to train wheat head segmentation models. Making use of a modified CycleGAN for synthesizing images allowed us to computationally generate large amounts of computationaly annotated images that closely resemble real wheat field images. While we showcased the utility of the proposed approach for wheat head segmentation, the proposed methodology can be applied to other crops and address the challenge of tedious and expensive manual annotation that is often a bottleneck for the full utilization of deep learning approaches for many applications. Our findings highlight the potential of utilizing synthetic data for developing deep learning models in domains where data annotation is tedious or costly.

Author Contributions: Conceptualization, F.M. and K.O.; GAN-based synthesization, model development, and evaluation, J.M.; software, J.M. and K.N.; resources, F.M. and K.O.; non-GAN data synthesization, K.N.; writing—original draft preparation, J.M. and K.O.; writing—review and editing, J.M., F.M., K.N., and K.O; visualization, J.M. and K.N.; supervision, F.M. and K.O.; project administration, K.O.; funding acquisition, K.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant number RGPIN-2023-04245. We also acknowledge the AI Research School at the University of Calgary, supported by Google's exploreCSR program, for providing the training opportunity to the students leading the deep learning model development for this research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were utilized in this study. These data can be found here: https://www.cs.usask.ca/ftp/pub/whs/ (accessed on 1 June 2023). The code used to generate synthetic data presented in this study are available on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GAN	Generative Adversarial Network
GWHD	Global Wheat Head Detection
CycleGAN	Cycle-Consistent Generative Adversarial Networks

References

1. Hao, S.; Zhou, Y.; Guo, Y. A Brief Survey on Semantic Segmentation with Deep Learning. *Neurocomputing* **2020**, *406*, 302–321. [CrossRef]
2. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
3. Richter, S.R.; Vineet, V.; Roth, S.; Koltun, V. Playing for Data: Ground Truth from Computer Games. In *European Conference on Computer Vision (ECCV), Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; LNCS; Springer International Publishing: Cham, Switzerland, 2016; Volume 9906, pp. 102–118.
4. Mardanisamani, S.; Maleki, F.; Hosseinzadeh Kassani, S.; Rajapaksa, S.; Duddu, H.; Wang, M.; Shirtliffe, S.; Ryu, S.; Josuttes, A.; Zhang, T.; et al. Crop lodging prediction from UAV-acquired images of wheat and canola using a DCNN augmented with handcrafted texture features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019.
5. Jia, B.; He, H.; Ma, F.; Diao, M.; Jiang, G.; Zheng, Z.; Cui, J.; Fan, H. Use of a Digital Camera to Monitor the Growth and Nitrogen Status of Cotton. *Sci. World J.* **2014**, *2014*, 602647. [CrossRef] [PubMed]
6. Xiong, Y.; Liang, L.; Wang, L.; She, J.; Wu, M. Identification of cash crop diseases using automatic image segmentation algorithm and deep learning with expanded dataset. *Comput. Electron. Agric.* **2020**, *177*, 105712. [CrossRef]
7. Behmann, J.; Mahlein, A.K.; Rumpf, T.; Römer, C.; Plümer, L. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. *Precis. Agric.* **2015**, *16*, 239–260. [CrossRef]

8. Al-Tamimi, N.; Langan, P.; Bernád, V.; Walsh, J.; Mangina, E.; Negrao, S. Capturing crop adaptation to abiotic stress using image-based technologies. *Open Biol.* **2022**, *12*, 210353. [CrossRef] [PubMed]
9. David, E.; Serouart, M.; Smith, D.; Madec, S.; Velumani, K.; Liu, S.; Wang, X.; Pinto, F.; Shafiee, S.; Tahir, I.S.; et al. Global wheat head detection 2021: An improved dataset for benchmarking wheat head detection methods. *Plant Phenomics* **2021**, *2021*, 1–9. [CrossRef]
10. Fourati, F.; Mseddi, W.S.; Attia, R. Wheat head detection using deep, semi-supervised and ensemble learning. *Can. J. Remote Sens.* **2021**, *47*, 198–208. [CrossRef]
11. Najafian, K.; Ghanbari, A.; Stavness, I.; Jin, L.; Shirdel, G.H.; Maleki, F. A Semi-Self-Supervised learning approach for wheat head detection using extremely small number of labeled samples. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 1342–1351.
12. Najafian, K.; Ghanbari, A.; Sabet Kish, M.; Eramian, M.; Shirdel, G.H.; Stavness, I.; Jin, L.; Maleki, F. Semi-Self-Supervised Learning for Semantic Segmentation in Images with Dense Patterns. *Plant Phenomics* **2023**, *5*, 0025. [CrossRef] [PubMed]
13. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
14. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
15. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976. [CrossRef]
16. Tian, Y.; Yang, G.; Wang, Z.; Li, E.; Liang, Z. Detection of apple lesions in orchards based on deep learning methods of cyclegan and yolov3-dense. *J. Sens.* **2019**, *2019*. [CrossRef]
17. Li, Y.; Zhan, X.; Liu, S.; Lu, H.; Jiang, R.; Guo, W.; Chapman, S.; Ge, Y.; Solan, B.; Ding, Y.; et al. Self-supervised plant phenotyping by combining domain adaptation with 3D plant model simulations: Application to wheat leaf counting at seedling stage. *Plant Phenomics* **2023**, *5*, 0041. [CrossRef]
18. Panda, S.S.; Ames, D.P.; Panigrahi, S. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sens.* **2010**, *2*, 673–696. [CrossRef]
19. Rawat, S.; L Chandra, A.; Desai, S.V.; Balasubramanian, V.; Ninomiya, S.; Guo, W. How Useful is Image-Based Active Learning for Plant Organ Segmentation? *Plant Phenomics* **2022**, *2022*, 1–11. [CrossRef]
20. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
21. Lee, D.H. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In Proceedings of the ICML 2013 Workshop: Challenges in Representation Learning (WREPL), Atlanta, GA, USA, 16–21 June 2013.
22. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations (ICLR). *arXiv* **2015**, arXiv:1412.6980
23. Bottou, L.; Curtis, F.E.; Nocedal, J. Optimization methods for large-scale machine learning. *Siam Rev.* **2018**, *60*, 223–311. [CrossRef]
24. Torbunov, D.; Huang, Y.; Yu, H.; Huang, J.; Yoo, S.; Lin, M.; Viren, B.; Ren, Y. UVCGAN: UNet Vision Transformer Cycle-consistent GAN for unpaired image-to-image translation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2023; pp. 702–712.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.