A Semi-self-supervised Learning Approach for Wheat Head Detection using Extremely Small Number of Labeled Samples

Keyhan Najafian[1,2], Alireza Ghanbari[3], Ian Stavness[2], Lingling Jin[2], Gholam Hassan Shirdel[3], and Farhad Maleki*[1]

[1]Augmented Intelligence & Precision Health Laboratory, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada
[2]Department of Computer Science, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
[3]Mathematics Department, Faculty of Sciences, University of Qom, Qom, Iran

Abstract

Most of the success of deep learning is owed to supervised learning, where a large-scale annotated dataset is used for model training. However, developing such datasets is challenging. In this paper, we develop a semi-self-supervised learning approach for wheat head detection. The proposed method utilized a few short video clips and only one annotated image from each video clip of wheat fields to simulate a large computationally annotated dataset used for model building. Considering the domain gap between the simulated and real images, we applied two domain adaptation steps to alleviate the challenge of distributional shift. The resulting model achieved high performance when applied to real unannotated datasets. When fine-tuned on the dataset from the Global Wheat Head Detection Challenge, the performance was further improved. The model achieved a mean average precision of 0.827, where an overlap of 50% or more between a predicted bounding box and ground truth was considered as a correct prediction. Although the utility of the proposed methodology was shown by applying it to wheat head detection, the proposed method is not limited to this application and could be used for other domains, such as detecting different crop types, alleviating the barrier of lack of large-scale annotated datasets in those domains.

1. Introduction

Considering the continuous growth of the human population, the use of computational approaches for increasing the quantity and quality of crops is imperative to ensure food security. These methods could be used for tasks such as predicting and monitoring crop growth, water stress, lodging, soil fertility, crop diseases, and deciding on effective prevention and management strategies for damage control [32, 21, 12].

Deep learning models have been successfully used to tackle real-world problems, including the vision for self-driving cars [38], speech recognition [20], and recommendation systems [7]. Deep learning models have also shown their potential for precision agriculture [13]. However, most of this success is owing to deep supervised learning models that rely on large-scale human-annotated datasets for model training. Human annotation is tedious, expensive, and time-consuming. This hinders developing deep learning models for many areas; therefore, methodologies that facilitate developing deep learning models in the absence of large-scale annotated datasets are of great interest.

Self-supervised learning is a branch of machine learning aimed at removing this barrier by relying on automatically generated labels for training deep learning models. In self-supervised learning, the supervisory signals come from data itself and are generated computationally, alleviating the need for human annotation. Often, a pretext task is designed, and a deep learning model is developed to learn the pretext task. Examples of pretext tasks are image colorization [39], image inpainting [23], and jigsaw puzzle [22]. To solve this pretext task, the model needs to extract latent representations of the inputs. These representations then can be used for accomplishing downstream tasks by fine-tuning the resulting model using a relatively small amount of training data.

Semi-supervised learning is another approach for dealing with the lack of large-scale labeled datasets. In semi-supervised learning, a small set of labeled data and a large

*Corresponding author (farhad.maleki@mail.mcgill.ca)

set of unlabeled data are used for model building. The goal is to develop a model with a performance superior to the model developed using only the labeled subset of data in a supervised manner or only the unlabeled subset of data in an unsupervised manner [4].

The need for large-scale annotated datasets still is a challenge faced by many application domains and impedes the development of deep learning models. In this paper, we propose a semi-self-supervised learning approach for wheat head detection to alleviate the need for large-scale training datasets. The contribution of this work is twofold: (1) We propose a methodology for object detection tasks that could lead to a high-performance model for wheat head detection. Although we focus on wheat head detection, the proposed method is independent of crop type and could be generalized to other crops. (2) The proposed method could also be used to facilitate and accelerate the labeling process, enabling the development of a wide range of supervised learning object detection systems.

2. Related works

Object detection models can be classified into two categories: one-stage detectors and two-stage detectors. One-stage object detection models use a single network that could be optimized end-to-end. OverFeat [29] is the first deep learning-based object detection model that followed this approach. SSD [16] and YOLO [24] are more recent methods with improved performance compared to OverFeat. Also, YOLO has gone through further improvements to increase its accuracy in object detection [25, 26, 2]. In two-stage detectors, first, a sparse set of proposals are generated. This step aims to filter out the majority of negative object location proposals while preserving the location proposals for actual objects. The goal of the second stage is to classify the remaining proposal to object classes or a background class. Selective Search [36] is one of the early works following this approach. R-CNN followed the same approach but used a convolutional neural network for the second stage [9]. R-CNN substantially improved performance compared to Selective Search. In this line of research, "Fast R-CNN" accelerated R-CNN by feeding the whole image once to the convolutional neural network instead of feeding each object proposal [8]. "Faster R-CNN" [27] further accelerated the "Fast R-CNN" by utilizing a feature proposal network instead of the Selective Search used in R-CNN and "Fast R-CNN".

Recently, Tan et al. [34] proposed EfficientDet, utilizing a weighted bi-directional feature pyramid network and a compound scaling approach [33]. They achieved the state-of-the-art despite using a relatively smaller network with a shorter inference time than previous methods.

A number of computer vision studies have reported approaches for detecting, localizing, and counting wheat heads from field images. The majority of these studies have employed supervised deep learning methods, such as customized image patch-based classification networks [41, 28] or standard object detection networks (e.g., Faster-RCNN) [18]. The advantage of detection methods is that individual wheat instances are localized, which may be important in certain applications such as measuring wheat head disease [31]. Obtaining a count or density of wheat heads may be sufficient for certain applications, e.g., informing selections in wheat breeding programs, in which density estimation approaches have shown promising results [37]. To avoid the cost of generating bounding box annotations for wheat heads, a recent work has explored unsupervised learning methods for wheat head counting [35]. The majority of previous wheat head detection/counting methods have been trained and evaluated on small-sized datasets, usually from one field, growing season, or growing region, which often result in models that are not generalizable. To address the potential lack of generalizability of previous wheat head counting approaches, the Global Wheat Head Dataset (GWHD) was assembled to increase the size and diversity of wheat head images available to researchers [5]. Indeed, these data create an interesting case of domain shift between wheat plants and images across the world and have been integrated as a sub-dataset within a meta-dataset of in-the-wild distribution shifts [14]. Recent work has begun to investigate domain adaptation methods for the wheat and plant counting [1, 10].

3. Method

Acquiring unlabeled image datasets for plants is often neither challenging nor expensive. However, the proposed pipeline further facilitates this by using short video clips of plant fields and background scenes, e.g. fields with no crop, and extracting image frames from these video clips to build a large unlabeled dataset. Top-down views of both background and wheat field video clips were taken using Samsung cameras with 12 and 48 Megapixels resolution. Figure 1 illustrates sample image frames from both the background and the wheat fields videos.

For each clip F_i of a wheat field, an image R_{F_i} is chosen, and the set of all wheat heads $H_{R_{F_i}} = \{h_j \mid 1 \leq j \leq n_{R_{F_i}}\}$ in that image are contoured, where $n_{R_{F_i}}$ is the number of wheat heads in the representative image R_{F_i}. Ideally, R_{F_i} should be a good representative of image frames in F_i. The main consideration is that the chosen frame includes wheat heads. The rest of the frames, excluding those in a buffer of 5 seconds before and after the chosen image R_{F_i}, are assigned to a set $\mathbb{I}_{F_i}$ of unlabeled image frames. The exclusion criterion is applied to avoid having images with overlap with R_{F_i}. For each video clip of background B_k, all image frames are extracted and added to a set $\mathbb{I}_{B_k}$ of background images, i.e., images with no wheat head.

Figure 1. Snapshots of video clips from the background scenes (rows 1–6) and the wheat fields (row 7).

To simulate an image/label pair, first, we randomly select a background image from a set $\mathbb{I}_{B_k}$ and a random subset of wheat heads from $H_{R_{F_i}}$. Then the chosen wheat heads are augmented and randomly placed on the background image. Through this process, the location of each wheat head is used to provide its corresponding bounding box annotation. It should be mentioned that we segmented R_{F_i}, instead of annotating it with bounding boxes. After rotating a bounding box, the axis-aligned box encompassing the rotated version might not be the tightest bounding box for the wheat head. This introduces error to the computational annotations. Therefore, we created a segmentation mask for R_{F_i} since the tightest bounding box for the rotated segment can be precisely calculated. The background images also undergo a data augmentation before the random placement of wheat heads on them. Through this process, for each background video clip B_K and each wheat field video clip F_i, we can simulate a set $\mathbb{S}_{k,i}$ of computationally annotated images. Figure 2 shows the process for simulating image/label pairs. This process could be used to simulate a large-scale computationally annotated dataset.

After simulating a large-scale annotated dataset, we train an object detection model for wheat heads. Since the data from the source domain (i.e., simulated images) is likely to have a distributional shift from the target domain (i.e., images of real wheat fields), we need to apply a domain adaptation approach.

We apply a two-stage domain adaptation approach. First, we create a dataset $\mathbb{D}$ including all 360 different rotations of each image R_{F_i} (rotations of $\theta = 0, \ldots, 359$ degrees). This ensures that all rotated versions of each image R_{F_i} are used. These images, unlike the simulated images, represent the wheat field and have smaller deviations from the real data distribution. To increase the variability of the images in $\mathbb{D}$, we used a pipeline of image augmentation, including a wide range of augmentations (see the Appendix). Figure 3 illustrates this process. Note that while the augmented images of wheat heads in R_{F_i} has been used for data simulation, the model still is not fully exposed to R_{F_i} itself. We fine-tune the model pretrained on the simulated dataset using $\mathbb{D}$.

As the second step for domain adaptation, the resulting fine-tuned model is used to detect objects in images of $\mathbb{I} = \cup_{\mathbb{I}_{F_i}}$, i.e., all unlabeled images extracted from different video clips of wheat fields. These pseudo-labels are then used as training data to further fine-tune the model using the real data.

3.1. Model training and evaluation

For all experiments, we used an implementation[1] of YOLO architecture [2] with the binary cross-entropy loss for objectness and CIoU loss [40] for bounding box regression. We used SGD optimizer [11] with a learning rate of 0.01and a momentum of 0.937. For training the model using the simulated dataset, we utilized 90% of samples from the simulated dataset for training and 10% for validation. For model evaluation, we used an external evaluation [19] using the test subset of the GWHD 2021 dataset [5], including 1381 annotated images.

We used 17 short background video clips ($B_1, \ldots, B_{17}$: a total of roughly 41 minutes) and three video clips of wheat fields (F_1, F_2, and F_3: a total of roughly 11 minutes). These resulted in the extraction of 294,511 image frames from background videos; three representative images of wheat fields (R_{F_1}, R_{F_2}, and R_{F_3}); and 17,979 image frames from the video clips of wheat fields, which were used as unlabeled datasets $\mathbb{I}_{F_1}$, $\mathbb{I}_{F_2}$, and $\mathbb{I}_{F_3}$. We simulated 36,000 computationally annotated samples. Figure 4 depicts examples of the simulated images.

Using all rotations of the images representing the three wheat fields, we created a dataset $\mathbb{D}$ of 1080 images. Figure 5 illustrates examples of images from the labeled dataset $\mathbb{D}$ with strong augmentations being applied. These images are originated from the three representative images extracted from the three video clips of wheat fields. Note that

[1] cloned from https://github.com/ultralytics/yolov5 on May 14

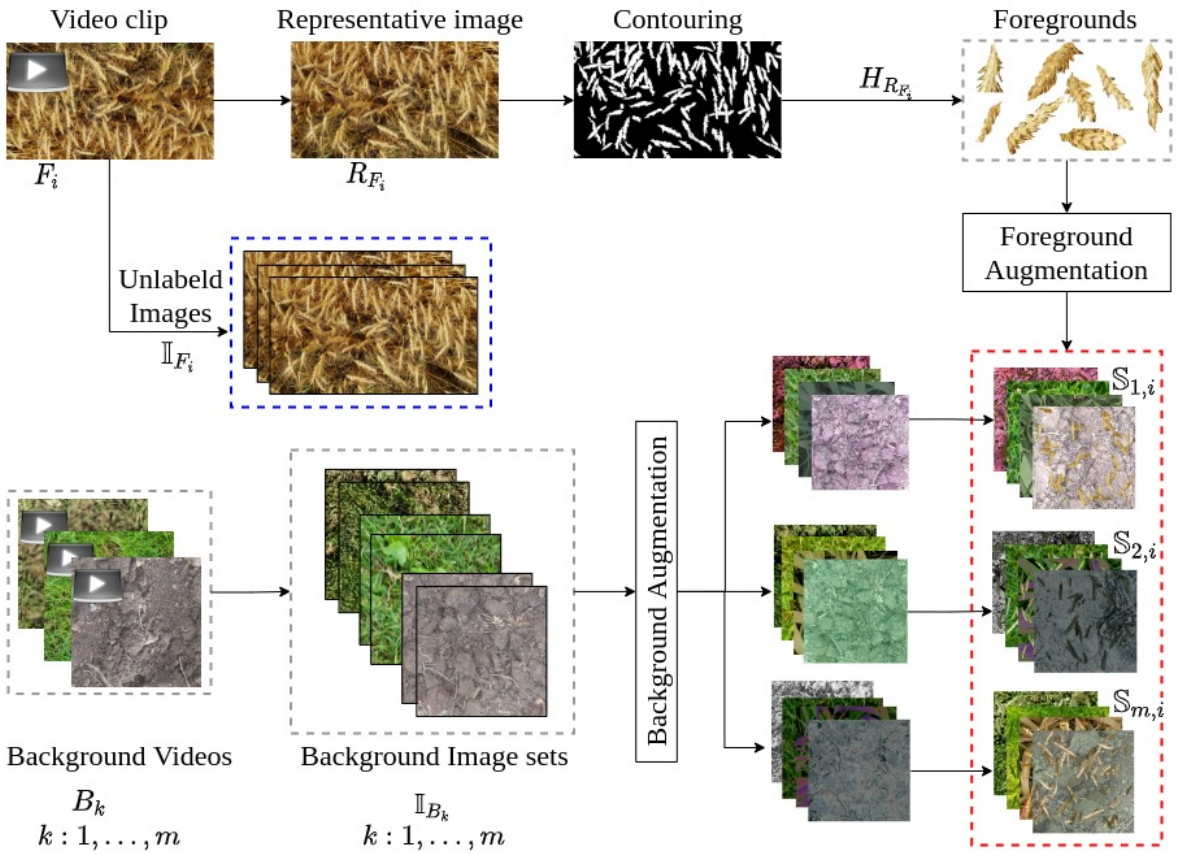

Figure 2. The procedure for simulating computationally annotated images.

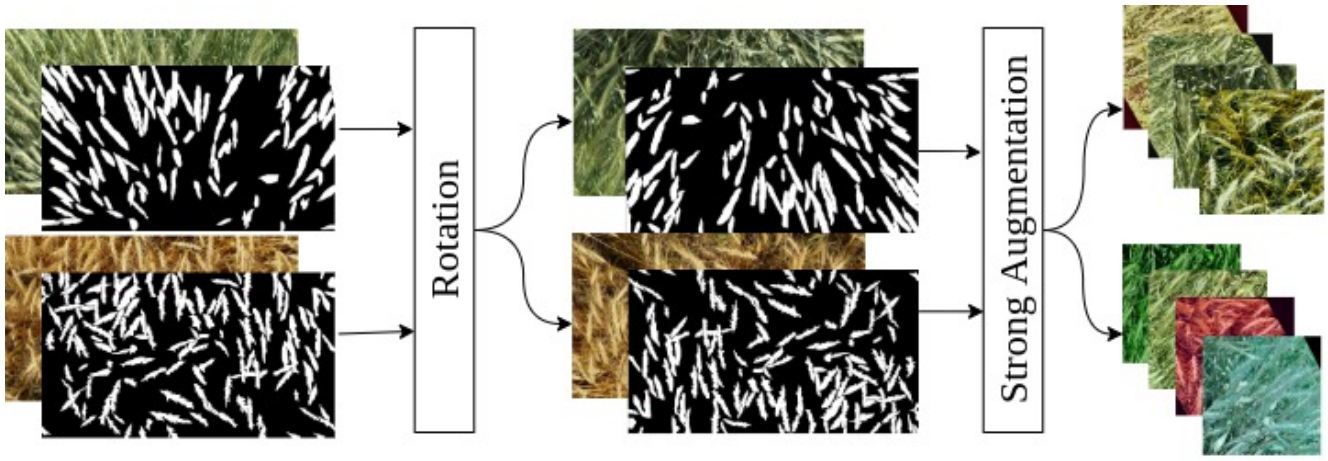

Figure 3. Developing a strongly augmented dataset using a few labeled images.

the strong augmentation is applied dynamically. In each epoch of the training process, a stochastic sequence of image augmentations is applied to each image resulting in an increased data variability across iterations and reducing the chance for overfitting.

The whole pipeline was implemented using Python version 3.9.4 and Pytorch version 1.8.1 on a NVIDIA TESLA V100 GPU machine. All augmentations were conducted using the Albumentation package version 0.5.2 [3]. Moreover, we used images of size 1024x1024 for all experiments to be consistent with the image sizes in the GWHD dataset.

Figure 4. Examples of simulated images. The distributional shift from real images of wheat fields can be observed.

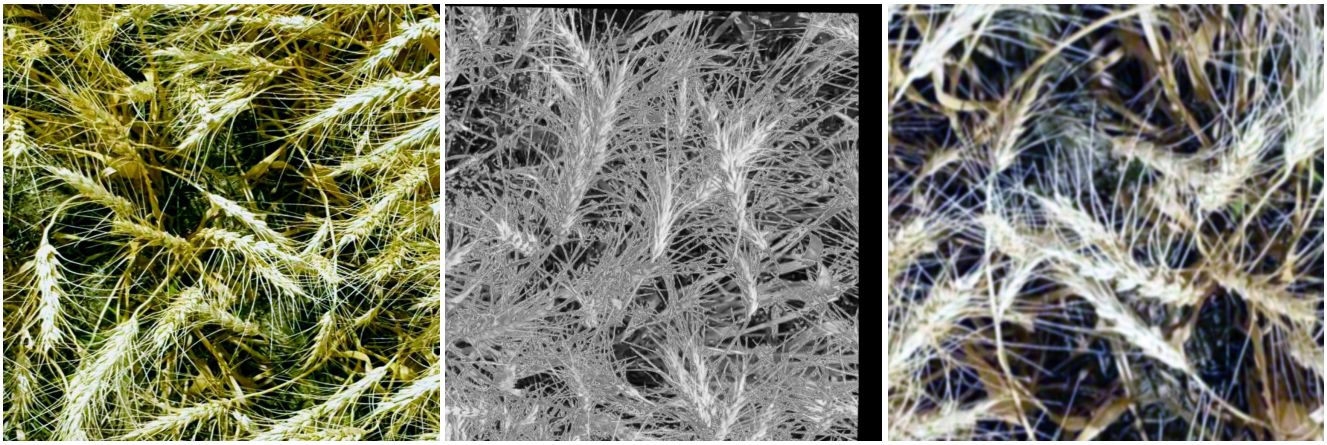

Figure 5. Examples of the strongly augmented images from the small labeled dataset.

4. Results

For each experiment, the YOLO model was trained for 25 epochs. Note that instead of simulating a smaller dataset and increasing the number of epochs, we chose to simulate a larger dataset and train the model for a smaller number of epochs. This was done to decrease the chance of overfitting to the simulated dataset.

Figures 6, 7, and 8 illustrate the performance of the trained models on randomly chosen images from the GWHD dataset (external evaluation). Table 1 shows the results of the external evaluation of models A, B, C, D, E, and a baseline model. Model A was the model trained on simulated images, and models B and C were the models resulting from the first and second steps of domain adaptation, respectively. Model D was the result of fine-tuning model C on the training subset of the GWHD dataset. Model E resulted from first pseudo-labeling the test set of the GWHD dataset using model D and then fine-tuning model D using the training set and the pseudo-labeled test set of the GWHD dataset. Note that models A, B, and C are not exposed to any information from the GWHD dataset. Model D is not exposed to the test set of the GWHD dataset, and model E is only exposed to the pseudo-labels for the test set of the GWHD dataset, not the actual labels. The baseline model was trained using the "train" set of the GWHD dataset.

Model C, i.e., the model that was fine-tuned using the pseudo-labels of the images frame dataset $\mathbb{I}$, achieved a mean average precision of 0.601. We further trained this model by first training on the training set of the GWHD dataset and then pseudo-labeling the test subset of the GWHD dataset. Then the resulting model was trained using all samples in the GWHD dataset.

5. Discussion

In this paper, we presented a semi-self-supervised learning approach followed by two domain adaptation steps for wheat head detection. The main contribution of this paper is the proposed semi-self-supervised approach followed by its domain adaptation steps, making it possible to pro-

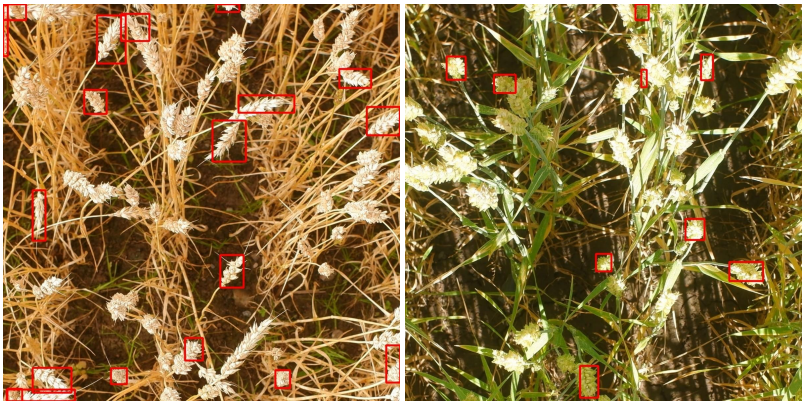

Figure 6. Examples of predictions made by the model A, which is only trained on the simulated dataset. The images are from the GWHD dataset unseen by the model. Due to the domain gap between the simulated data and the real data, model A failed to detect many wheat heads.

Figure 7. Example predictions made by model B, i.e. the model resulting from the first step of domain adaptation. We fine-tuned model A (see Figure 6) on the strongly augmented labeled dataset $\mathbb{D}$. The resulting model is referred to as Model B. The images are from the GWHD dataset unseen by the model.

Figure 8. Example predictions made by model C, i.e., the model resulting from the second step of domain adaptation. We use model B (see Figure 7) to pseudo-label the large dataset of unlabeled images frames. Then model B is further fine-tuned using the images in $\mathbb{I}$ and their pseudo-labels. The fine-tuned model is referred to as model C. The images are from the GWHD dataset unseen by the model.

duce a high-performance wheat head detection model with short unannotated video clips and only three contoured images. The model developed using the simulated dataset achieved a mean average precision of 0.09 on the test set of the GWHD dataset, highlighting the domain gap and the need for domain adaptation. The domain adaptation steps

Table 1. The performance of the trained models in terms of precision, recall, and mean average precision for five models. Model A was the model trained using the simulated dataset. Model B was the model resulting from fine-tuning model A by training on dataset $\mathbb{D}$. Model C was the model resulting from fine-tuning model B using the pseudo-labels of the unlabeled dataset $\mathbb{I}$ extracted from the video clips of wheat fields. The pseudo-labels are generated using model B. Model D was generated by further training of model C on the training subset of the GWHD dataset. We used model D to pseudo-label images in the test set of the GWHD dataset. Then all the images in the GWHD dataset were used to further train model D. This resulted in the final model (model E). The baseline model was trained using the "train" set of the GWHD dataset in a supervised manner. All the performance measures were calculated on the test set of the GWHD dataset.

Model	Precision	Recall	mAP 50
A	0.318	0.130	0.090
B	0.779	0.477	0.513
C	0.801	0.534	0.601
D	0.867	0.731	0.794
E	0.877	0.774	0.827
Baseline	0.832	0.688	0.741

resulted in a substantial improvement resulting in a model with a mean average precision of 0.601. When fine-tuned on the dataset from the Global Wheat Head Detection Challenge, the performance was further improved. The model achieved a mean average precision of 0.827, where an overlap of 50% or more between a predicted bounding box and ground truth was considered as a correct prediction. This is substantially higher than the baseline model trained using the GWHD dataset, which is a large-scale annotated dataset.

In self-supervised learning, a pretext task is often designed to computationally determine data labels without manual annotation. Then, using these computationally generated labels, a model is trained in a supervised manner to learn the pretext task, with the goal of capturing a latent representation for input data. This model is retrained on a relatively small dataset for a downstream task of interest. Note that the computationally generated labels might be completely irrelevant to the downstream task. In this paper, we used a semi-self-supervised approach followed by two domain adaptation steps. Our semi-self-supervised approach computationally generates labeled data for the main task. This is different from self-supervised learning, in which a pretext task such as image rotation or inpainting is used to generate labels that might be entirely irrelevant to the main task. The semi-self-supervised approach also differs from self-supervised learning as it requires a few manually labeled samples to synthesize a large computationally annotated dataset. It also differs from semi-supervised learning, in which a small set of labeled data and a relatively larger set of unlabeled data are used for model development. We only used a few manually labeled samples. Using background images with no wheat head, we computationally generated a large dataset for the main task (wheat head detection here). Note that the semi-self-supervised learning approach should not be mistaken with the second step of domain adaptation, i.e., the pseudo-labeling approach, which is a semi-supervised approach.

We used a cut and paste approach similar to Dwibedi et al. [6] for generating simulated images. Dwibedi et al. utilized a cut and paste approach to synthesize images for developing an object instance augmentation. They first trained an FCN network [17] to be able to separate foreground objects. The trained model was used to segment the objects in the Big Berkeley Instance Recognition Dataset [30]. These segmented objects were then placed on background images from the UW Scenes dataset [15]. In this paper, we used a similar cut and paste approach for data simulation. However, unlike the proposed method by Dwibedi et al. that relies on another model to be trained and used for segmenting foreground objects to simulate images, we only used three annotated images and used video clips both for background and wheat fields. This makes our approach easily adaptable to different domains where a large-scale dataset is not available. In addition, video clips of crops could be easily acquired at little to no cost. In addition, we utilized domain adaptation techniques that substantially improved the performance of the model trained on the simulated datasets.

We used video clips for developing deep learning models for wheat head detection. The use of video clips as the data source for crop detection has several benefits. Collecting video clips has little to no costs in comparison to developing large-scale datasets. Also, it takes a few minutes to take a video clip while developing a large-scale image dataset is tedious and time-consuming. This approach also makes it possible to extract a large number of images from videos computationally. Therefore, our approach makes developing deep learning models for crop detection more accessible for domains where large-scale datasets are unavailable.

We utilized an external model evaluation approach using the test set of the GWHD dataset. Note that we differentiate the validation and evaluation performance. The former is calculated on the validation set and the latter on the test set. External evaluation is the most reliable means of model evaluation, providing an unbiased estimate of the generalization error [19]. Validation error, on the other hand, is a biased estimate of generalization error [19]. The models developed using the semi-self-supervised approach and the two steps of domain adaptation are not exposed to the data from the GWHD dataset either in the semi-self-supervised phase or in the domain adaptation steps. Therefore, the resulting performance measures are highly reliable considering the scale of the GWHD dataset and its high degree of variability, representing different stages of wheat growth. It should be mentioned that our evaluation for the models

fine-tuned on the GWHD dataset (i.e., models D and E) is considered internal evaluation [19], as we used the test set of the same dataset for evaluation.

In this paper, we only used three short clips of wheat fields representing three stages of wheat growth. We expect that including video clips representing different growth stages of wheat further improves the model performance. Also, we only used one representative image for each field. This was conducted to highlight the feasibility of object detection using very few annotated images of wheat fields. Annotating more images from each video clip could result in a higher diversity in the training data, more representative samples, and further improvement in the model performance.

We used an implementation of YOLO; however, the proposed method is independent of the model being used for object detection. Also, we used only one representative image for each wheat field, and this image was partially used for building both training and validation sets (but not the test set). Therefore, the training and validation (not the test data) might be partially dependent. Note that the proposed method alleviates this issue following two different strategies. In data simulation for the semi-self-supervised approach, we used background images independent of the representative images; also, the wheat heads extracted from a representative image were augmented and randomly placed on the background images. In the first step of domain adaptation, we used strong data augmentation to decorrelate the images used for fine-tuning (training) and validation as much as possible. By using only two representative images for each wheat field, this issue would be resolved, and the performance measure for model validation would be more reliable.

We observed that pseudo-labeling is an effective way of improving model performance during which a model is exposed to the data coming from the target data distribution. If appropriately utilized, pseudo-labeling could be a practical approach for domain adaptation. In this paper, we used only one step of pseudo-labeling. This could be further improved by applying two steps of pseudo-labeling.

6. Conclusion

In this research, we proposed a semi-self-supervised approach followed by two domain adaptation steps for wheat head detection. The proposed approach only uses short video clips of wheat fields and background scenes. Using only a few contoured images and the video clips, we simulated a large-scale computationally annotated dataset and a large-scale unlabeled dataset. Using these datasets, the proposed approach led to a high-performing model. The model was further improved when fine-tuned on the GWHD dataset. Although we showed the utility of the proposed approach for wheat head detection, it is not limited to this purpose. The proposed method could be used for a wide range of applications, including detection for other crop types.

Appendix

In this research, we used a wide range of image augmentations from the Albumentations package. These was acomplished through developing an stochastic sequence of the following image augmentations: *Blur*, *ChannelShuffle*, *CLAHE*, *ColorJitter*, *Equalize*, *FancyPCA*, *Flip*, *GaussianBlur*, *GaussNoise*, *GlassBlur*, *HorizontalFlip*, *HueSaturationValue*, *InvertImg*, *MedianBlur*, *MultiplicativeNoise*, *Posterize*, *RandomBrightnessContrast*, *RandomFog*, *RandomGamma*, *RandomRain*, *RandomSnow*, *RandomSunFlare*, *RGBShift*, *Solarize*, *ToGray*, *VerticalFlip*.

References

[1] Tewodros W Ayalew, Jordan R Ubbens, and Ian Stavness. Unsupervised domain adaptation for plant organ counting. In *European Conference on Computer Vision*, pages 330–346. Springer, 2020.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[3] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020.

[4] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-supervised Learning*. Adaptive computation and machine learning. MIT Press, 2010.

[5] Etienne David, Simon Madec, Pouria Sadeghi-Tehran, Helge Aasen, Bangyou Zheng, Shouyang Liu, Norbert Kirchgessner, Goro Ishikawa, Koichi Nagasawa, Minhajul A. Badhon, Curtis Pozniak, Benoit de Solan, Andreas Hund, Scott C. Chapman, Frédéric Baret, Ian Stavness, and Wei Guo. Global wheat head detection (GWHD) dataset: A large and diverse dataset of high-resolution RGB-Labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics*, 2020:3521852, Aug 2020.

[6] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1301–1310, 2017.

[7] Zeshan Fayyaz, Mahsa Ebrahimian, Dina Nawara, Ahmed Ibrahim, and Rasha Kashef. Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *Applied Sciences*, 10(21):7748, 2020.

[8] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[10] Dario Gogoll, Philipp Lottes, Jan Weyler, Nik Petrinic, and Cyrill Stachniss. Unsupervised domain adaptation for transferring plant classification systems to new field environments, crops, and robots. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2636–2642. IEEE, 2020.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[12] Yoshio Inoue. Satellite-and drone-based remote sensing of crops and soils for smart farming–a review. *Soil Science and Plant Nutrition*, 66(6):798–810, 2020.

[13] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018.

[14] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[15] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *2011 IEEE International Conference on Robotics and Automation*, pages 1817–1824. IEEE, 2011.

[16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[18] Simon Madec, Xiuliang Jin, Hao Lu, Benoit De Solan, Shouyang Liu, Florent Duyme, Emmanuelle Heritier, and Frederic Baret. Ear density estimation from high resolution rgb imagery using deep learning technique. *Agricultural and Forest Meteorology*, 264:225–234, 2019.

[19] Farhad Maleki, Nikesh Muthukrishnan, Katie Ovens, Caroline Reinhold, and Reza Forghani. Machine learning algorithm validation: from essentials to advanced applications and implications for regulatory certification and deployment. *Neuroimaging Clinics*, 30(4):433–445, 2020.

[20] Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6):9411–9457, 2021.

[21] Sara Mardanisamani, Farhad Maleki, Sara Hosseinzadeh Kassani, Sajith Rajapaksa, Hema Duddu, Menglu Wang, Steve Shirtliffe, Seungbum Ryu, Anique Josuttes, Ti Zhang, 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 Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[22] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[23] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[24] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[25] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7263–7271, 2017.

[26] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28:91–99, 2015.

[28] Pouria Sadeghi-Tehran, Nicolas Virlet, Eva M Ampe, Piet Reyns, and Malcolm J Hawkesford. DeepCount: in-field automatic quantification of wheat spikes using simple linear iterative clustering and deep convolutional neural networks. *Frontiers in Plant Science*, 10:1176, 2019.

[29] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. OverFeat: Integrated recognition, localization and detection using convolutional networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[30] Arjun Singh, James Sha, Karthik S Narayan, Tudor Achim, and Pieter Abbeel. BigBIRD: A large-scale 3D database of object instances. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 509–516. IEEE, 2014.

[31] Wen-Hao Su, Jiajing Zhang, Ce Yang, Rae Page, Tamas Szinyei, Cory D Hirsch, and Brian J Steffenson. Automatic evaluation of wheat resistance to fusarium head blight using dual mask-RCNN deep learning frameworks in computer vision. *Remote Sensing*, 13(1):26, 2021.

[32] Babak Talebpour, Ufuk Türker, and Uğur Yegül. The role of precision agriculture in the promotion of food security. *International Journal of Agricultural and Food Research*, 4(1), 2015.

[33] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[34] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2020.

[35] Jordan R Ubbens, Tewodros W Ayalew, Steve Shirtliffe, Anique Josuttes, Curtis Pozniak, and Ian Stavness. AutoCount: Unsupervised segmentation and counting of organs in field images. In *European Conference on Computer Vision*, pages 391–399. Springer, 2020.

[36] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[37] Haipeng Xiong, Zhiguo Cao, Hao Lu, Simon Madec, Liang Liu, and Chunhua Shen. TasselNetv2: in-field counting of wheat spikes with context-augmented local regression networks. *Plant Methods*, 15(1):1–14, 2019.

[38] A Yoganandhan, SD Subhash, J Hebinson Jothi, and V Mohanavel. Fundamentals and development of self-driving cars. *Materials Today: Proceedings*, 33:3303–3310, 2020.

[39] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016.

[40] Zhaohui Zheng, Ping Wang, Dongwei Ren, Wei Liu, Rongguang Ye, Qinghua Hu, and Wangmeng Zuo. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Transactions on Cybernetics*, 2021.

[41] Chengquan Zhou, Dong Liang, Xiaodong Yang, Hao Yang, Jibo Yue, and Guijun Yang. Wheat ears counting in field conditions based on multi-feature optimization and TWSVM. *Frontiers in Plant Science*, 9:1024, 2018.