Improving Wheat Head Detection: A Data-Centric Approach by Domain Variance Reduction[*]

**** ****
****
****, ****
****@****.***

**** ****
****
****, ****
****@****.***

ABSTRACT

AI-assisted agriculture helps in obtaining accurate data about a worldwide staple food – wheat. It is vital to estimate the density and size of the wheat head (spike on top of the plant with the grain) from images or manual survey. Such estimates help to make informed decisions related to the health and maturity of the wheat plants. Computer vision has been employed recently to detect wheat heads. It is a challenging task as wheat heads are small and vary in color depending on their growth phase. Also, there is significant variance among the images due to field conditions and planting patterns. This paper presents a data-driven approach to reduce variance among wheat head images to improve detection results. We use preprocessing steps to standardize the color histogram of images and train a YOLOv5 model to detect the wheat heads. We use histogram matching to make a color histogram of every image consistent with the global average color histogram of the training set. This is followed by contrast limited adaptive histogram equalization (CLAHE) to improve the contrast on a local scale and improve the visibility of the wheat heads. With 0.715 Average Domain Accuracy (ADA), our approach achieves state-of-the-art (SOTA) results, outperforming the existing SOTA (0.700 ADA) approach by 1.5% while requiring almost 5x fewer training computation steps.

CCS CONCEPTS

• **Computing methodologies** → **Object detection**; *Image processing*; • **Applied computing** → *Agriculture*;

KEYWORDS

Domain generalization, histogram matching, object detection, wheat head, yolov5

ACM Reference Format:
**** **** and **** ****. 2022. Improving Wheat Head Detection: A Data-Centric Approach by Domain Variance Reduction. In *Proceedings of ACM SAC Conference (SAC'22).* ACM, New York, NY, USA, Article 4, 7 pages. https://doi.org/xx.xxx/xxx_x

1 INTRODUCTION

Due to the recent advancements in the field of computer vision(CV) and artificial intelligence(AI), tasks like classification [15, 21, 31, 37, 39], semantic segmentation [4, 12, 35] and object detection [2, 11, 27] are becoming increasingly efficient and accurate. It has led to the automation of laborious and monotonous tasks. This has benefited agriculture, which is now rapidly changing due to the infusion of AI and CV in traditional farming methods. Also, due to proliferation of Internet of Things(IoT) devices, agriculture data [5, 14, 17, 20, 22, 28] is now available on a large scale, facilitating data driven AI.

Recently, automated detection of wheat heads has been of particular interest in the research community as wheat is one of the most planted food grains in the world. It is consumed universally due to its rich nutritional value, including protein and fiber. The United Nations Food and Agriculture Organization(FAO) estimated the global wheat production to be around 765.76 million tons in 2019 [9]. It highlights the importance of wheat production in the global food supply chain. Thus, it is crucial to identify growth and predict wheat yield to ensure abundant production. Wheat head counting is an essential factor in determining the yield of the wheat crop. However, traditional methods rely on manual labor to count the wheat heads. Such methods are expensive, time consuming, prone to error and extremely inefficient in the modern high volume production scenario. According to an estimate by [29], manual wheat head counting has about 10 % error rate.

Recent deep neural network-based object detection methods can be used to detect wheat heads. These methods do not require manual feature generation. They can generate highly generalized detection models, which can count the wheat heads in a fraction of the time compared to manual counting. The solution presented in [44] used a custom implementation of the YOLOv4 [2] object detection model and trained on the Global Wheat Challenge(GWC) 2020 dataset [7] to achieve a mAP of 94.0%. Gong et al. [13] introduces Spatial Pyramid Pooling at the head and tail in the backbone of YOLOv4 [2] and achieve a mAP of 94.5% as compared to standard YOLOv4 with 91.4%. Li et al. [24] used RetinaNet [26] and transfer learning along with image enhancement techniques to detect wheat heads and also provided a comparison with a Faster-RCNN [34] based approach. Wang et al. [42] proposed an EfficientDet [40] and a BiFPN based approach to address occlusion robust wheat head detection. Misra et al.[30] use a UNet style segmentation-based approach to identify wheat heads followed by patch cleaning and counting which is provided as online service to count wheat heads. Fourati et al. [10] focus on a semi-supervised learning (pseudo-labeling), test time

[*] Produces the permission block, and copyright information

SAC'22, April 25 –April 29, 2022, Brno, Czech Republic
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-8713-2/22/04...$15.00
https://doi.org/xx.xxx/xxx_x

augmentations and post processing applied over ensemble of Faster-RCNN models (multiple resolutions). Lightweight models for real-time wheat head detection from UAV images have been proposed in [16, 19, 45], whereas Khakhi et al. [19] used the MobileNetv2 [36] model as backbone.

Even though the deep learning-based approaches for detecting wheat heads provide good results, they are far from being perfect. The primary issue faced by such approaches irrespective of the model is the lack of abundant and varied data. The wheat heads can vary in color, size, and shape depending on the growth phase. Factors like soil, time of day, wind also contribute to variance in the data. Detection of wheat heads in such varied conditions is challenging and requires a robust approach to address domain variance. Apart from this, it needs a balance between efficiency and detection accuracy.

This paper uses YOLOv5 [18], an existing SOTA method in object detection. It proposes a preprocessing step to reduce domain variance among the images. Instead of a new model, we present the impact of data preprocessing in challenging multi-domain data. Our approach achieves SOTA accuracy on the GWC 2021 dataset [8]. We present a comparison with the winners of GWC 2021, and the computational efficiency of our approach is compared with other models. The proposed approach uses YOLOv5 to quickly detect wheat heads and tackle domain variance. The following sections discuss the dataset, methods, and results of our study.

2 DATASET

2.1 Preliminaries

The GWHD 2021 [8] dataset we use in this study is an extension of GWHD 2020 [7] dataset, which had 193,634 wheat heads labeled in 4,793 images from seven countries. The updated GWHD 2021 dataset improves upon the previous dataset by reexamining and correcting labels and adding 1,722 images from five more countries, and 81,553 wheat heads. The GWHD 2021 dataset used in this study contains 6,515 images each with dimensions of 1024 x 1024 and 2,75,187 unique labeled wheat heads. The images are captured in individual sessions (domain), with a total of 47 sessions. A *domain* is a set of images acquired at the same location, during a coherent timestamp (usually a few hours), with a specific sensor, and corresponds to a particular development stage. Figure 1A represents the distribution of images across the 47 domains. The wheat can be from any of the four development stages in a particular session, namely: *Post-flowering, Filling, Filling-Ripening, Ripening*. The distribution of images in the development stages is given in Figure 1B. The variation between the different development stages is also shown in the example images.

The training dataset contains 3,655 images from 18 domains which are acquired from Europe and Canada. The test set contains 2,856 images from 29 domains acquired from North America (except Canada), Asia, Oceania, and Africa. The dataset is imbalanced, with images per session ranging from a minimum of 14 to a maximum of 747.

2.2 Challenges with the Dataset

Detection of wheat heads is a challenging task. Following are the factors which contribute to the difficulty in wheat head detection.

- A large number of objects are in the frame.
- Overlapping wheat heads.
- Variations in appearance based on genotype.
- Orientation of wheat heads.
- Different development stage of the wheat.

Unlike traditional object detection tasks, where only a few objects are present in an image, there are hundreds of wheat heads per image. The density of wheat heads also poses another challenge where wheat heads partially or fully overlap other wheat heads. The genotype of wheat determines the appearance of wheat heads and can vary based on the country of origin. The orientation of wheat heads is also a significant challenge since wheat heads cannot be easily identified if facing vertically up or down. The wheat heads are generally well visualized from the side-profile. The development stage of wheat determines the color and appearance of wheat heads, ranging from green to yellow. The average histogram of Red-Green-Blue(RGB) colors can be visualized in Figure 2. The histogram corresponding to the training and test sets without preprocessing is presented in Figure 2A. It can be seen that the average color histogram of the test set is different from the training set.

3 PROPOSED METHOD

3.1 Domain Variance Reduction

The most differentiating factors in images from different domains are color variations due to different genotypes, development stages, sensor type, time of the day, or imaging conditions. This variation results in wheat heads and surrounding leaves being different in color across domains, as shown in Figure 1B.

In order to reduce domain variance, we introduce two-step preprocessing. As the first step of preprocessing, we perform histogram matching [3] on all the training and test images. We calculate the global average histogram of the training set (Figure 2) and match the histogram of each images to the global average and blend the resultant image with the original image in a 0.5:0.5 ratio. This step reduces the variation in color hues among the domains. This step is followed by CLAHE [32] with a clip_limit of 1.5. CLAHE is an adaptive algorithm and takes local information into account while equalizing the histogram. Thus, it produces a more realistic version of contrast enhancement when compared to normal histogram equalization which uses the global context of the image. The second step ensures sufficient contrast between wheat heads and the rest of the image and enhances the visibility of the wheat heads. Figure 2B shows the results of the preprocessing step on the average color histogram of training and test sets. As observed, the histogram distribution of the training and test set is consistent after the preprocessing step.

Since we are using a semi-supervised training approach along with augmentation which introduces variation in color, it is important to make the actual training data and pseudo labeled test data have a similar underlying visual style. It prevents unrealistic outlier images after the image augmentation steps. In order to present the quantitative analysis of the preprocessing step, we present the mean and standard deviation of average values RGB colors across the entire training and test set before and after the preprocessing step in Table 1. The variation in the average value of RGB color across images is reduced which is confirmed by lowering of the

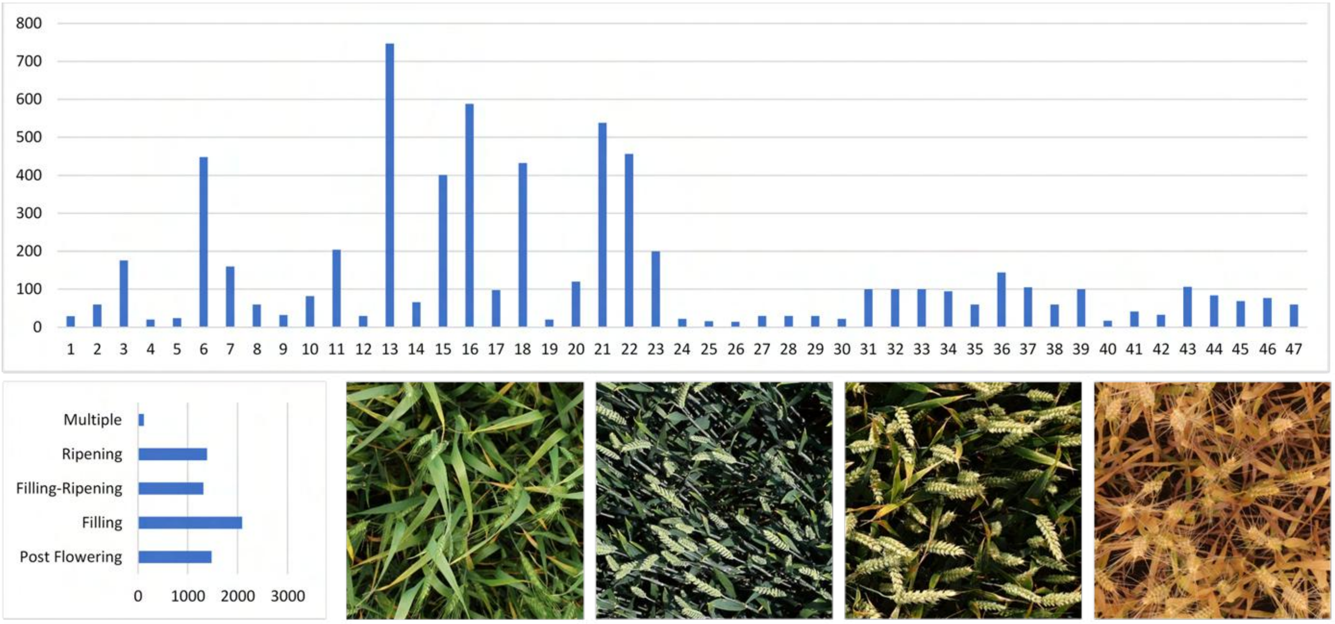

Figure 1: (A: Top) The distribution of Global Wheat Head Detection(GWHD) [6] images with the 47 domains. (B: Bottom-left) Distribution of images for the growth stages. (B: Bottom-right) Example images from (left to right) Post-Flowering, Filling, Filling-Ripening and Ripening stages

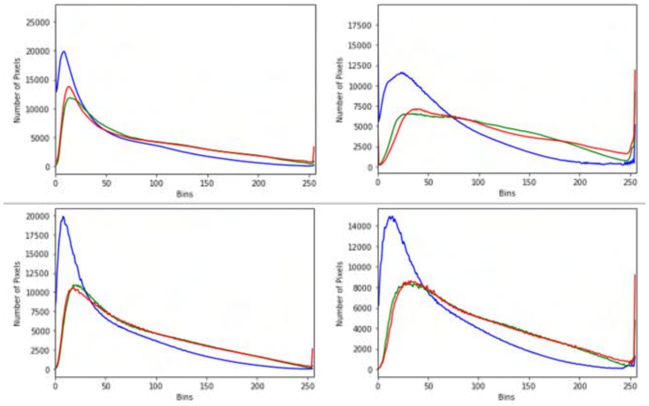

Figure 2: The global average color histogram for training (left) and test (right) set before (A: Top) and after (B: Bottom) the preprocessing step.

standard deviation of average color values after the preprocessing step. The resultant images after the preprocessing step are presented in Figure 3. As it can be seen, the visibility of the wheat heads is increased after the preprocessing step.

3.2 Augmentations

The data is limited considering 6000 images and factoring in the variation across domains. As shown in Figure 1, some domains are marginally represented and are prone to underfitting. In order to improve the representation and introduce geometrical variations,

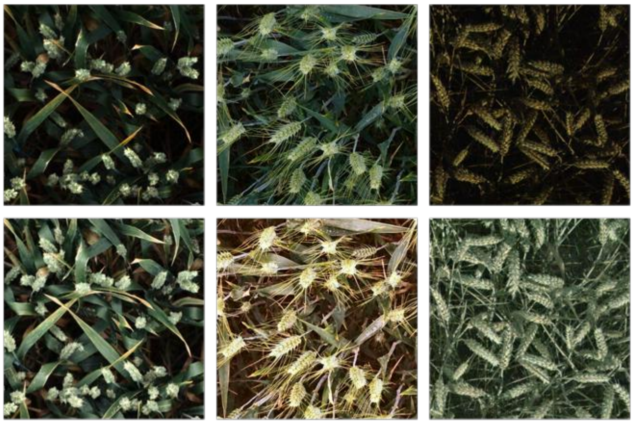

Figure 3: Example images before (Top) and after (Bottom) the preprocessing step.

Table 1: Mean and Standard Deviation (SD) of average color values across traning and test set. (NI: Normal Image, PI: Preprocessed Image).

		Red		Green		Blue	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Train	**NI**	82.91	32.87	82.18	26.39	54.19	17.54
	PI	98.82	20.62	98.31	15.80	70.52	10.89
Test	**NI**	109.43	26.22	104.28	20.82	64.26	21.20
	PI	111.03	18.19	108.46	13.55	74.56	11.73

Table 2: List of augmentations used to train the model with their hyperparameters.

Augmentation	Hyperparameter Value
Rotate	0.15 (angle)
Translate	0.1 (fraction of image size)
Shear	0.1 (angle)
Scale	0.9 (fraction of image size)
Flip (horizontal/vertical)	0.5 (probability)
Mosaic	1 (probability)
HSV	H: 0.015, S: 0.7, V: 0.4

we use data augmentation. Data augmentation and the operating hyperparameters are presented in Table 2.

Geometric transforms like rotate, translate, shear, scale, and flip introduce more variations in the shape, size, and orientation of the wheat heads. Mosaic [43] uses the collage of multiple images to introduce color variations in images and thus helps the model be robust across different domains. The HSV augmentation on preprocessed images is valuable as the underlying images have a similar distribution. Thus, the application of such augmentation does not create an unrealistic image. Mosaic is a novel data augmentation technique that uses multiple images to create a collage used to train the model instead of a single image. This data augmentation introduces high variation in a single image, making the model robust to change in the domain. All the augmentation is provided by the YOLOv5 model library [18].

3.3 Object Detection Models

YOLOv5: In this approach, we use the YOLOv5 architecture [18] with the yolov5x6 configuration for training on high resolution 1024x1024 images. The backbone is based on CSPDarknet, which incorporates Cross Stage Partial Network (CSPNet) [41] into Darknet [33]. The CSPnet helps reduce parameters and increases inference speed as compared to similar large-scale deep learning backbones. The neck of YOLOv5 is Pyramid Attention Network (PANet) [23] which follows the Feature Pyramid Network (FPN) [25] style of structure which helps in the propagation of low-level features. When combined with adaptive feature pooling, it introduces attention. It identifies helpful information in the feature layers to ensure the flow of essential features from low-level feature maps to high-level feature maps. Finally, the head of YOLOv5 consists of four convolution layers that detect objects of four different sizes. The size of wheat heads varies from very small to very large, depending on the mode of the image. Thus multi scale object detection ensures wheat heads are detected irrespective of the relative size in an image.

Reference Model (FasterRCNN): To validate the applicability of our preprocessing step, we also use the reference model provided by GWC 2021 [8] for comparison. It is a FasterRCNN [34] based object detection model with a Resnet50 [15] backbone. It has the same parameters as described in GWHD 2020 paper [7]. This model is a two-stage object detection model. But, despite having 1/3$^{\text{rd}}$ of trainable parameter as compared to YOLOv5, training the model takes nearly 4x time.

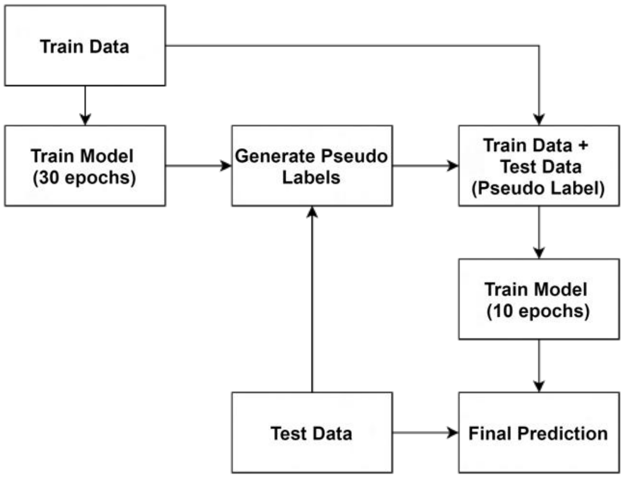

Figure 4: Overview of training routine.

3.4 Training

In Figure 4, we present the training routine used to train the YOLOv5 model. We train the model in two steps: 1) Fully supervised and 2) Semi supervised using pseudo labels. We train our model to training set for 30 epochs. The trained model predicts labels for the test set and generates a pseudo label dataset for semi-supervised training. The combined (training and pseudo labeled) test set are used to train the model for another ten epochs. The model trained on the pseudo labeled data predicts the final output on the test set. The model is trained using an Stochastic Gradient Descent(SGD) optimizer with an initial learning rate of 0.01 and Binary Cross Entropy Loss (BCELoss) for object classification. The Intersection over Union (IoU) value 0.2 is used during training. All the experiments are performed on a Google Colab environment with 24 GB RAM and 16 GB vRAM Nvidia P100 GPU.

For the reference model, we follow similar training routine. However, we use the hyperparameter configuration provided by the GWHD 2020 paper [7] and as recommended train the model for one epoch each during supervised and semi-supervised training.

4 RESULTS

The models are trained and tested on only GWHD 2021 [8]. The testing of the models has been performed on the official evaluation platform [1], thus ensuring fair and one-to-one comparison without any evaluation algorithm bias. We present the metric used to quantify the performance, comparison with the existing SOTA solutions, ablation on the use of preprocessing step, and the efficiency comparison among approaches.

4.1 Metrics

The metrics defined by GWC 2021 [8] for evaluation of predicted bounding boxes is ADA. Accuracy for each image is calculated as:

$$AI = \frac{TP}{TP + FN + FP} \qquad (1)$$

Table 3: Ablation on preprocessing step. (ADA: Average domain accuracy, NI: Normal Image, PI: Preprocessed Image)

	Trained on PI		**Trained on NI**	
	Test PI	**Test NI**	**Test PI**	**Test NI**
Proposed (ADA)	0.709	0.715	0.683	0.698

where: TP is True Positive when a ground truth bounding box is matched with any one predicted box. FP is False Positive when a predicted bounding box does not match any ground truth box. FN is False Negative when a ground truth bounding box matches no predicted box. Bounding box matching: Two boxes (ground truth and predicted) are considered matched if their Intersection over Union is higher than a threshold of 0.5.

Average Domain Accuracy: The GWC 2020 identified issues with the global average metric. It enabled domains with higher image count to dominate the average result and thus is not representative of the domain level results. In order to address this, GWC 2021 proposed ADA as the metric for evaluation. ADA ensures that results from each domain are given equal weightage and are not overwhelmed by an over performing domain. The average domain accuracy for final evaluation is calculated as follows:

$$ADA = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{n_d} \times \sum_{i=1}^{n_d} AI_{di} \tag{2}$$

Where D represents the total number of domains (47 in GWHD 2021), n_d represents the number of images belonging to domain d, $d_i$represents the i^{th} image in the domain d and AI represents the accuracy for an image.

Special cases: In cases where there is no bounding box in the ground truth, and at least one box is predicted, accuracy is equal to 0. If there is no bounding box in both ground truth and prediction, then accuracy is equal to 1.

4.2 Ablation on Preprocessing Step

We perform ablation by training the model on preprocessed images(PI) and normal images(NI). The trained models are then tested using PI(s) and NI(s), and the results are presented in Table 3. The model trained on PI(s) consistently outperforms the model trained on NI(s). We also perform similar ablation on the reference model to verify the trend in results. It is evident from Table 3 that the model trained using the proposed preprocessing step gives better results when tested on NI(s). The test set contains a higher number of filling-ripening and ripening images where the wheat heads are highly pronounced. The performance improves due to the improved color representation and appearance of wheat heads in the PI test set.

4.3 Comparison with State-of-the-Art

Table 4: Comparison with top ranking methods of GWC 2021 [8].

Solution	**ADA**
Proposed approach	**0.715**
1st rank	0.700
2nd rank	0.695
3rd rank	0.695
Reference or Baseline	0.492

Table 5: Comparison of efficiency metrics with top ranking methods of GWC 2021 [8].

Approach	**#Models**	**#Epochs**	**TTA**	**Post Processing**
Proposed approach	1	40	No	Non-max suppression
1st rank	4 (Ensemble)	230	Yes	Weighted Box Fusion
2nd rank	1	600	Yes	Weighted Box Fusion
3rd rank	1	300	Yes	None

We compare our model with the existing SOTA solutions for GWC 2021 [8]. Table 4 shows a comparison of our model with the winners of GWC 2021. Our model with 0.715 ADA achieves nearly 1.5% improvement over the winning solution with 0.700 ADA. We also provide results of the reference solution presented in [7]. The reference solution uses a relatively simple model(Faster-RCNN) without any ensemble approach, does not use any advanced data augmentation techniques, or post processing and thus has a poor performance (0.492 ADA).

4.4 Computational Efficiency

The computational efficiency of our approach is the crucial differentiating factor over the existing approaches. Table 5 presents the efficiency comparison metrics of our approach with the winners of GWC 2021. The comparison takes into account the number of models trained, number of epochs for training, use of test time augmentation and post processing. Our approach uses a single model as compared to the four model ensemble used by the 1st rank solution of GWC 2021. Our model trains in fewer epochs, nearly 5x less than the first and 14x less than the second-place solution. The third place solution uses a custom Dynamic Color Transform Network (DCTN) which is trained in conjunction with the object detection network. The DCTN learns to improve the color of the image and applies it during inference. Both top two winning solutions also use test time augmentation (TTA) and Weighted box fusion [38] to improve the prediction results. Adding TTA increases inference time by nearly 2x to 3x. Our proposed approach does not use TTA or Weighted box fusion which are computationally expensive. The proposed approach only uses Non-max suppression as a post-processing step and yet achieves SOTA results.

In Figure 5, we present the two sets of results, for correct detection and the cases with detection failure. Figure 5A presents cases where the model correctly estimates the bounding boxes of all the wheat heads. Figure 5B shows the failure cases where the model incorrectly predicts the bounding box with no wheat heads present. Specifically, the center image in Figure 5B presents a case where non-max suppression fails to remove overlapping bounding boxes. The false-positive prediction was mainly observed in cases like the

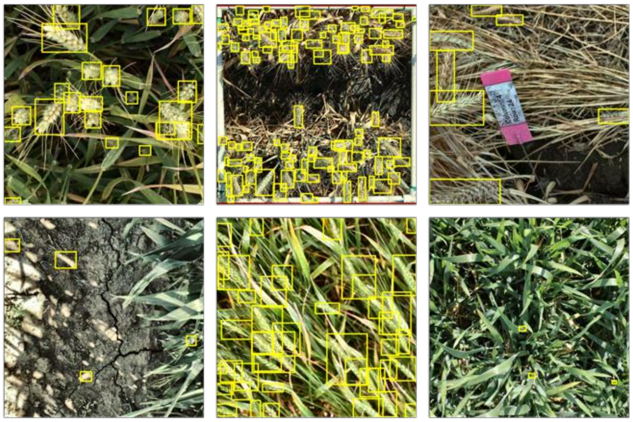

Figure 5: (A: Top) The cases of correct bounding box prediction. (B: Bottom) The cases of false positive or incorrect bounding box predictions.

first image in Figure 5B. Sunlight projects bright patches on the ground, similar to wheat heads causing false detection.

5 CONCLUSION

In this paper, we have presented a data-driven approach to wheat head detection in a multi-domain scenario. We use the YOLOv5 model for wheat head detection and train it using the semi-supervised (pseudo label) approach. The main highlight of this paper is the proposed preprocessing step to reduce domain variance among the images. It makes the color histogram consistent with the global average value, thereby reducing the disparity between over-represented and under-represented domains. Our proposed approach achieves SOTA results. We also present the validity of the proposed preprocessing step by performing ablation against the model trained on normal images without any preprocessing. We also provide a computational efficiency comparison of our approach with the existing SOTA approaches. Our proposed approach without any ensemble approach, test time augmentation, and expensive post-processing can predict accurate bounding boxes even when trained for limited epochs. Thus, the proposed approach can improve the results of models irrespective of their complexity and drive the adoption of data-driven approaches.

A HEADINGS IN APPENDICES

The rules about hierarchical headings discussed above for the body of the article are different in the appendices. In the **appendix** environment, the command **section** is used to indicate the start of each Appendix, with alphabetic order designation (i.e., the first is A, the second B, etc.) and a title (if you include one). So, if you need hierarchical structure *within* an Appendix, start with **subsection** as the highest level. Here is an outline of the body of this document in Appendix-appropriate form:

A.1 Introduction

A.2 Dataset

A.2.1 Preliminaries.

A.2.2 Challenges with the Dataset.

A.3 Proposed Method

A.3.1 Domain Variance Reduction.

A.3.2 Augmentations.

A.3.3 Object Detection Models.

A.3.4 Training.

A.4 Results

A.4.1 Metrics.

A.4.2 Ablation on Preprocessing Step.

A.4.3 Comparison with State-of-the-Art.

A.4.4 Computational Efficiency.

A.5 References

ACKNOWLEDGMENTS

The authors convey their thanks to the reviewers for providing valuable comments on the manuscript.

REFERENCES

[1] GWC 2021 AIcrowd. 2021. Global Wheat Challenge. Retrieved September 4, 2021 from https://www.aicrowd.com/challenges/global-wheat-challenge-2021/submissions

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. 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. 2020. Albumentations: fast and flexible image augmentations. *Information* 11, 2 (2020), 125.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

[5] Jeffrey A Cruz, Xi Yin, Xiaoming Liu, Saif M Imran, Daniel D Morris, David M Kramer, and Jin Chen. 2016. Multi-modality imagery database for plant phenotyping. *Machine Vision and Applications* 27, 5 (2016), 735–749.

[6] Etienne David, Simon Madec, Pouria Sadeghi-Tehran, Helge Aasen, Bangyou Zheng, Shouyang Liu, Norbert Kirchgessner, Goro Ishikawa, Koichi Nagasawa, Minhajul A Badhon, et al. 2020. 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 (2020).

[7] Etienne David, Franklin Ogidi, Wei Guo, Frederic Baret, and Ian Stavness. 2021. Global Wheat Challenge 2020: Analysis of the competition design and winning models. *arXiv preprint arXiv:2105.06182* (2021).

[8] Etienne David, Mario Serouart, Daniel Smith, Simon Madec, Kaaviya Velumani, Shouyang Liu, Xu Wang, Francisco Pinto Espinosa, Shahameh Shafiee, Izzat SA Tahir, et al. 2021. Global Wheat Head Dataset 2021: an update to improve the benchmarking wheat head localization with more diversity. *arXiv preprint arXiv:2105.07660* (2021).

[9] Food and Agriculture Organization of the United Nations. 2019. Crops and Livestock Data. Retrieved September 4, 2021 from http://www.fao.org/faostat/en/#data/QCL

[10] Fares Fourati, Wided Souidene Mseddi, and Rabah Attia. 2021. Wheat Head Detection using Deep, Semi-Supervised and Ensemble Learning. *Canadian Journal of Remote Sensing* (2021), 1–13.

[11] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[13] Bo Gong, Daji Ergu, Ying Cai, and Bo Ma. 2021. Real-time detection for wheat head applying deep neural network. *Sensors* 21, 1 (2021), 191.

[14] Wei Guo, Bangyou Zheng, Andries B Potgieter, Julien Diot, Kakeru Watanabe, Koji Noshita, David R Jordan, Xuemin Wang, James Watson, Seishi Ninomiya,

et al. 2018. Aerial imagery analysis–quantifying appearance and number of sorghum heads for applications in breeding and agronomy. *Frontiers in plant science* 9 (2018), 1544.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[16] Ming-Xiang He, Peng Hao, and You-Zhi Xin. 2020. A Robust Method for Wheatear Detection Using UAV in Natural Scenes. *IEEE Access* 8 (2020), 189043–189053.

[17] David Hughes, Marcel Salathé, et al. 2015. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060* (2015).

[18] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg, wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham. 2021. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations. https://doi.org/10.5281/zenodo.4679653

[19] Saeed Khaki, Nima Safaei, Hieu Pham, and Lizhi Wang. 2021. Wheatnet: A lightweight convolutional neural network for high-throughput image-based wheat head detection and counting. *arXiv preprint arXiv:2103.09408* (2021).

[20] David LeBauer et al. 2020. Data from: TERRA-REF, an open reference data set from high resolution genomics, phenomics, and imaging sensors. *Dryad Dataset* (2020).

[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[22] Simon Leminen Madsen, Solvejg Kopp Mathiassen, Mads Dyrmann, Morten Stigaard Laursen, Laura-Carlota Paz, and Rasmus Nyholm Jørgensen. 2020. Open plant phenotype database of common weeds in denmark. *Remote Sensing* 12, 8 (2020), 1246.

[23] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. 2018. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180* (2018).

[24] Jingbo Li, Changchun Li, Shuaipeng Fei, Chunyan Ma, Weinan Chen, Fan Ding, Yilin Wang, Yacong Li, Jinjin Shi, and Zhen Xiao. 2021. Wheat ear recognition based on RetinaNet and transfer learning. *Sensors* 21, 14 (2021), 4845.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature Pyramid Networks for Object Detection. arXiv:cs.CV/1612.03144

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[28] Hao Lu, Zhiguo Cao, Yang Xiao, Bohan Zhuang, and Chunhua Shen. 2017. TasselNet: counting maize tassels in the wild via local counts regression network. *Plant methods* 13, 1 (2017), 1–17.

[29] Simon Madec, Xiuliang Jin, Hao Lu, Benoit De Solan, Shouyang Liu, Florent Duyme, Emmanuelle Heritier, and Frederic Baret. 2019. Ear density estimation from high resolution RGB imagery using deep learning technique. *Agricultural and forest meteorology* 264 (2019), 225–234.

[30] Tanuj Misra, Alka Arora, Sudeep Marwaha, Ranjeet Ranjan Jha, Mrinmoy Ray, Rajni Jain, AR Rao, Eldho Varghese, Shailendra Kumar, Sudhir Kumar, et al. 2021. Web-SpikeSegNet: deep learning framework for recognition and counting of spikes from visual images of wheat plants. *IEEE Access* 9 (2021), 76235–76247.

[31] Aditya Parikh, Mehul S Raval, Chandrasinh Parmar, and Sanjay Chaudhary. 2016. Disease detection and severity estimation in cotton plant from unconstrained images. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE, 594–601.

[32] Stephen M Pizer. 1990. Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen m. pizer, r. eugene johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group. In *Proceedings of the First Conference on Visualization in Biomedical Computing, Atlanta, Georgia*, Vol. 337.

[33] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015), 91–99.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[37] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[38] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. 2021. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing* 107 (2021), 104117.

[39] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

[40] Mingxing Tan, Ruoming Pang, and Quoc V Le. 2020. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10781–10790.

[41] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. 2020. CSPNet: A new backbone that can enhance learning capability of CNN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 390–391.

[42] Yiding Wang, Yuxin Qin, and Jiali Cui. 2021. Occlusion Robust Wheat Ear Counting Algorithm Based on Deep Learning. *Frontiers in Plant Science* 12 (2021), 1139.

[43] Zhiwei Wei, Chenzhen Duan, Xinghao Song, Ye Tian, and Hongpeng Wang. 2020. AMRNet: Chips Augmentation in Aerial Images Object Detection. *arXiv preprint arXiv:2009.07168* (2020).

[44] Baohua Yang, Zhiwei Gao, Yuan Gao, and Yue Zhu. 2021. Rapid Detection and Counting of Wheat Ears in the Field Using YOLOv4 with Attention Module. *Agronomy* 11, 6 (2021), 1202.

[45] Jianqing Zhao, Xiaohu Zhang, Jiawei Yan, Xiaolei Qiu, Xia Yao, Yongchao Tian, Yan Zhu, and Weixing Cao. 2021. A Wheat Spike Detection Method in UAV Images Based on Improved YOLOv5. *Remote Sensing* 13, 16 (2021), 3095.