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

Etienne David, Simon Madec, Pouria Sadeghi-Tehran, Helge Aasen, Bangyou Zheng, Shouyang Liu, Norbert Kirchgessner, Goro Ishikawa, Koichi Nagasawa, Minhajul Arifin Badhon, et al.

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.

E. David[1,2] †, S. Madec[1,2] †, P. Sadeghi-Tehran[3], H. Aasen[4], B. Zheng[5], S. Liu[2,6], N. Kirchgessner[4], G. Ishikawa[7], K. Nagasawa[8], M.A. Badhon[9], C. Pozniak[10], B. de Solan[1], A. Hund[4], S.C. Chapman[5,11], F. Baret[2,6], I. Stavness[9*], W. Guo[12*]

[1] Arvalis, Institut du végétal, 3, rue Joseph et Marie Hackin, 75116 Paris, France
[2] UMR1114 EMMAH, INRAE, Centre PACA, Bâtiment Climat, Domaine Saint-Paul, 228 route de l'Aérodrome, CS 40509, 84914 Avignon CEDEX, France
[3] Plant Sciences Department, Rothamsted Research, Harpenden, United Kingdom
[4] Institute of Agricultural Sciences, ETH Zurich, Universitätstrasse 2, 8092 Zurich, Switzerland
[5] CSIRO Agriculture and Food, Queensland Biosciences Precinct 306 Carmody Road, St Lucia, 4067, QLD, Australia
[6] Plant Phenomics Research Center, Nanjing Agricultural University, Nanjing, China
[7] Institute of Crop Science, National Agriculture and Food Research Organization, Japan
[8] Hokkaido Agricultural Research Center, National Agriculture and Food Research Organization, Japan
[9] Department of Computer Science, University of Saskatchewan, Canada
[10] Department of Plant Sciences, University of Saskatchewan, Canada
[11] School of Food and Agricultural Sciences, The University of Queensland, Gatton 4343, QLD,Australia
[12] Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 midori-cho, nishi-tokyo city, Tokyo, Japan
*Corresponding author. Email: ian.stavness@usask.ca ; guowei@g.ecc.u-tokyo.ac.jp
†These authors contributed equally to this work

Detection of wheat heads is an important task allowing to estimate pertinent traits including head population density and head characteristics such as sanitary state, size, maturity stage and the presence of awns. Several studies developed methods for wheat head detection from high-resolution RGB imagery. They are based on computer vision and machine learning and are generally calibrated and validated on limited datasets. However, variability in observational conditions, genotypic differences, development stages, head orientation represents a challenge in computer vision. Further, possible blurring due to motion or wind and overlap between heads for dense populations make this task even more complex. Through a joint international collaborative effort, we have built a large, diverse and well-labelled dataset, the Global Wheat Head detection (GWHD) dataset. It contains 4,700 high-resolution RGB images and 190,000 labelled wheat heads collected from several countries around the world at different growth stages with a wide range of genotypes. Guidelines for image acquisition, associating minimum metadata to respect FAIR principles and consistent head labelling methods are proposed when developing new head detection datasets. The GWHD is publicly available at http://www.global-wheat.com/ and aimed at developing and benchmarking methods for wheat head detection.

Keywords: dataset, wheat spike/ear/head, detection, deep learning, RGB

1. Introduction

Wheat is the most cultivated cereal crop in the world, along with rice and maize. Wheat breeding progress in the 50's was key for food security of emerging countries when Norman Borlaug developed semi-dwarf wheats and a complementary agronomy system (the Doubly Green Revolution), saving 300 million people from starvation [1]. However, after increasing rapidly for decades, the rate of increase in wheat yields has slowed down since the early 1990s [2], [3]. Traditional breeding still relies to a large degree on manual observation. Innovation to increase genetic gain may come from genomic selection, new high-throughput phenotyping techniques or a combination of both [4]–[7]. These techniques are key to select important wheat traits linked to yield potential, disease resistance or adaptation to abiotic stress. Even though high throughput phenotypic data acquisition is already a reality, developing efficient and robust models to extract traits from raw data remains a significant challenge. Among all traits, wheat head number per unit ground area is a major yield component and is still manually evaluated in breeding trials, which is labour intensive and leads to measurement errors around 10% [8], [9]. Thus developing a sensing-based methods to increase the throughput and accuracy of wheat heads counting in the field is highly desired to help breeders manipulating the balance between yield components (plant number, head density, grains per head, grain weight).

Thanks to increases in GPU performances and the emergence of large scale datasets [10], [11] , deep learning has become the state of the art for computer vision for object detection [12], instance segmentation [13], semantic segmentation [14] and image regression [15], [16]. Recently, several authors have proposed deep learning methods tailored to image-based plant phenotyping [17]–[19]. Several methods have been proposed for wheat head quantification from RGB high resolution images. In [8],[9], the authors demonstrated the potential to detect wheat heads with a Faster-RCNN object detection network. They estimated in [8] a relative counting error of around 10% for such methods when the image resolution is controlled. In [20], the authors developed an encoder-decoder CNN model for semantic segmentation of wheat heads which outperformed traditional handcrafted computer vision techniques. [21] developed a wheat head detection and probabilistic tracking model to characterize the motion of wheat plants grown in the field.

While previous studies have tested wheat head detection methods on individual datasets, in practice these deep learning models are difficult to scale to real-life phenotyping platforms, since they are trained on limited datasets, with expected difficulties when extrapolating to new situations [8], [22], [23]. Most training datasets are limited in terms of genotypes, geographic areas and observational conditions. Wheat head morphology may significantly differ between genotypes with notable variation in head morphology, including size, inclination, color and the presence of awns. The appearance of heads and the background canopy also change significantly from emergence to maturation due to ripening and senescence [24]. Further, planting densities and patterns vary globally across different cropping systems and environments, with possible overlap between heads for the higher head densities.

A common strategy for handling limited dataset is to train a CNN model on part of a phenotyping trial field and test it on the remaining fraction [25]. This is a fundamental flaw of empirical approaches against causal models: there is no theoretical guarantee that a CNN is robust on new acquisitions. In addition, comparison between methods from different authors require large datasets. Unfortunately, such large and diverse phenotyping head counting datasets do not exist today because they are mainly acquired independently by single institutions, limiting the number of genotypes, the environmental and the observational conditions used to train and test the models. Further, because the labelling process is burdensome and tedious, only a small fraction of the acquired images is processed. Finally, labelling protocols may be different between institutions, which will limit model performances when trained over shared labelled datasets.

To fill the need for a large and diverse wheat head dataset with consistent labelling, we developed the Global Wheat Head Detection (GWHD) dataset that can be used to benchmark methods proposed in the computer vision community. The GWHD dataset results from the harmonization of several datasets coming from nine different institutions across seven countries and three continents. This paper details the data collection, the harmonization process across image characteristics and labelling, the organization of a wheat head detection challenge. Finally, we discuss the issues raised while generating the dataset and propose guidelines for future contributors who wish to expand the GWHD dataset with their labelled images.

2. Dataset composition

2.1. Experiments

The labelled images come from datasets collected between 2016 and 2019 by nine institutions at ten different locations (Table 1) covering genotypes from Europe, North America, Australia and Asia. These individual datasets are called "sub-datasets". They were acquired over experiments following different growing practices, with row spacing varying from 12.5 cm (ETHZ_1) to 22 cm (UQ_1). The characteristics of the experiments are presented in Table 1. They include low sowing density (UQ_1, UTokyo_1, UTokyo_2), normal sowing density (Arvalis_1, Arvalis_2, Arvalis_3, INRAE_1, part of NAU_1) and high sowing density (RRes_1, ETHZ_1, part of NAU_1). The GWHG dataset covers a range of pedoclimatic conditions including very productive context such as the loamy soil of Picardy area in France (Arvalis_3), silt-clay soil in mountainous conditions like the Swiss Plateau (ETHZ_1) or Alpes de Haute Provence (Arvalis_1, Arvalis_2). In the case of Arvalis_1, Arvalis_2, UQ_1, NAU_1, the experiments were designed to compare irrigated and water stressed modalities.

Sub-dataset name	Institution	Country	Lat (°)	Long (°)	Year	Nb. of dates	Targeted stages	Row spacing (cm)	Sowing density (seeds·m-2)	Nb. of genotypes
UTokyo_1	NARO & UTokyo	Japan	36.0N	140.0E	2018	3	Post-flowering	15	186	66
UTokyo_2	NARO & UTokyo	Japan	42.8N	143.0	2016	6	Flowering*	12.5	200	1
Arvalis_1	Arvalis	France	43.7N	5.8E	2017	3	Post-flowering - Ripening	17.5	300	20
Arvalis_2	Arvalis	France	43.7N	5.8E	2019	1	Post-flowering	17.5	300	20
Arvalis_3	Arvalis	France	49.7N	3.0E	2019	3	Post-flowering - Ripening	17.5	300	4
INRAE_1	INRAE	France	43.5N	1.5E	2019	1	Post-flowering	16	300	7
USask_1	University of Saskatchewan	Canada	52.1N	106.W	2019	1	n.a	30.5	250	16
RRes_1	Rothamsted Research	UK	51.8N	0.36W	2016	1	n.a	n.a	350	6
ETHZ_1	ETHZ	Switzerland	47.4N	8.6E	2018	1	n.a	12.5	400	354
NAU_1	Nanjing Agric. University	China	31.6N	119.4E	2018	1	Flowering*	20	300 or 450	5
UQ_1	UQueensland	Australia	27.5S	152.3E	2016	1	Flowering - Ripening	22	150	8

Table 1: Characteristics of the experiments used to acquire images for GWHD Dataset.

* images were checked carefully to ensure that heads have fully developed and flowering.

2.2. Image acquisition

The GWHD dataset contains RGB images captured with a wide range of ground-based phenotyping platforms and cameras (Table 2). The height of the image acquisition ranges between 1.8 m and 3 m above the ground. The camera focal length varies from 10 to 50 mm with a range of sensor sizes. The differences in camera setup lead to a range of Ground Sampling Distance (GSD) ranging from 0.10 to 0.62 mm with the half field of view along the image diagonal varying from 10° to 46°. Assuming that wheat heads are 1.5 cm diameter, the acquired GSDs are high enough to visually detect heads and even awns. Although all images were acquired at nadir-viewing direction, some geometric distortions may be observed for few sub-datasets due to the different lens characteristics of the cameras used, excerpts of the acquired images have different. Datasets UTokyo_1 and ETHZ_1 are particularly affected by this issue. Each institution acquired images from different platforms, including handheld, cart, mini-vehicle and gantry systems. The diversity of camera sensors and acquisition configurations resulted in a wide range of image properties that will help training deep learning models that can better generalize across the diversity of image acquisition conditions.

Sub-dataset name	Vector	Camera	Focal length (mm)	Field of view (°) *	Shooting mode	Image size (pixels)	Distance to Ground (m)	GSD (mm/px)
UTokyo_1	cart	Canon PowerShot G9 X Mark II	10	38.15	automatic	5472×3648	1.8	0.43
UTokyo_2	handheld	Olympus µ850 & Sony DSC-HX90V	7/4	45.5	automatic	3264×2488 & 4608×3456	1.7	0.6
Arvalis_1	handheld	Sony Alpha ILCE-6000	50 & 60	7.1	automatic	6000×4000	2.9	0.10-0.16
Arvalis_2	handheld	Sony RX0	7.7	9.99	automatic	800x800†	1.8	0.56
Arvalis_3	handheld	Sony RX0	7.7	9.99	automatic	800x800†	1.8	0.56
INRAE_1	handheld	Sony RX0	7.7	9.99	automatic	800×800†	1.8	0.56
USask_1	mini-vehicle	FLIR Chameleon3 USB3	16	19.8	fixed	2448×2048	2	0.45
RRes_1	gantry	Prosilica GT 3300 Allied Vision	50	12.8	automatic	3296×2472	3-3.5§	0.33-0.385
ETHZ_1	gantry	Canon EOS 5D Mark II	35	32.2	fixed	5616×3744	3	0.55
NAU_1	handheld	Sony RX0	24	16.9	automatic	4800×3200	2	0.21
UQ_1	handheld	Canon 550D	55	17.3	automatic	5184×3456	2	0.2

Table 2: Images characteristics of the sub-datasets composing the GWHD Dataset. All cameras looked vertically downward.

* The field of view is measured diagonally. The reported measure is the half-angle.

† Original images were cropped, and a sub-image of size 800x800 was extracted from the central area

§ The camera was positioned perpendicular to the ground and automatically adjusts to ensure a 2.2 m distance was maintained between the camera and canopy.

2.3. Data harmonization

A selection of images was first conducted to ensure that they could be well interpreted. Images acquired at too early stages were removed when heads were not clearly visible (Figure 2, (d)). Images were also mostly acquired before the appearance of head senescence since heads tend to overlap when they start to bend at this stage.

Object scale, i.e. the size of the object in pixels, is important to design the detection methods [8]. Object scale depends on the size (mm) of the object and on the resolution of the image. Wheat head dimensions may vary across genotypes and growth conditions, with however around 1.5 cm diameter and 10 cm length. However, the actual image resolution at the head level was varying significantly between sub-datasets: the GSD varies by a factor of 5 (Table 1) while the actual resolution at the head level depends also on canopy height and the panoramic effect of the camera. The panoramic effect will be much larger when images were acquired too close to the canopy. Images were therefore rescaled to keep more similar resolution at the head level. Bilinear interpolation was used to up- or down-sample the original images. The scaling factor applied to each sub-dataset is displayed in Table 3.

Most deep-learning algorithms are running over individual squared patches. The size of the patches should be adjusted to limit the probability to get heads crossing the edges of the patches. Images were therefore split into 1024 x 1024 squared patches containing roughly 20 to 60 heads with only few heads crossing the edges. The number of patches per original images was varying between 1 to 6 (Table 3). These squared patches will be termed "images" in the following of this study.

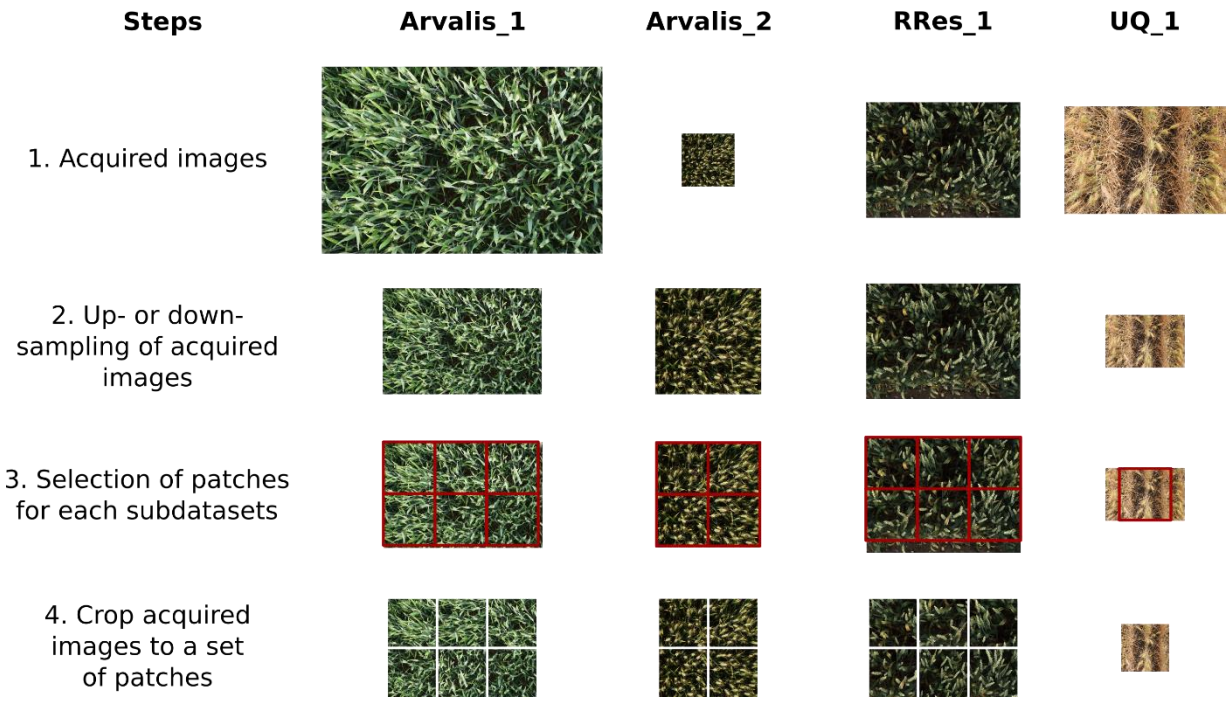

Figure 1: Overview of the harmonization process conducted. Images were first rescaled using bilinear interpolation up- or -down-sampling techniques. Then, the rescaled images were split into 1024 x 1024 squared patches.

2.4. Labelling

A web-based labelling platform was set up to handle the evaluation and labelling of the shared sub-datasets using the coco annotator (https://github.com/jsbroks/coco-annotator; [26]). The platform hosts all the tools required to label objects. In addition, it also grants simultaneous access to different users, thus allowing contributions from all institutions. Wheat heads were interactively labelled by drawing bounding boxes that contained all the pixels of the head. Labelling is difficult if heads are not clearly visible, i.e. if they are masked by leaves or other heads. We did not label partly hidden heads, unless at least one spikelet was visible. This was mostly the case for images acquired at an early stage when heads were not fully emerged. Overlap among heads was more frequently observed when the images were acquired using a camera with a wide field of view as in UTokyo_2 or ETHZ_1. These overlaps occurred mainly towards the borders of the images with a more oblique view angle. When the bounding box was too large to include the awns, it was restricted to the head only (Figure 2 (a)). Further, heads cropped at the image edges were labelled only if more than 30% of their basal part was visible (Figure 2, (e)).

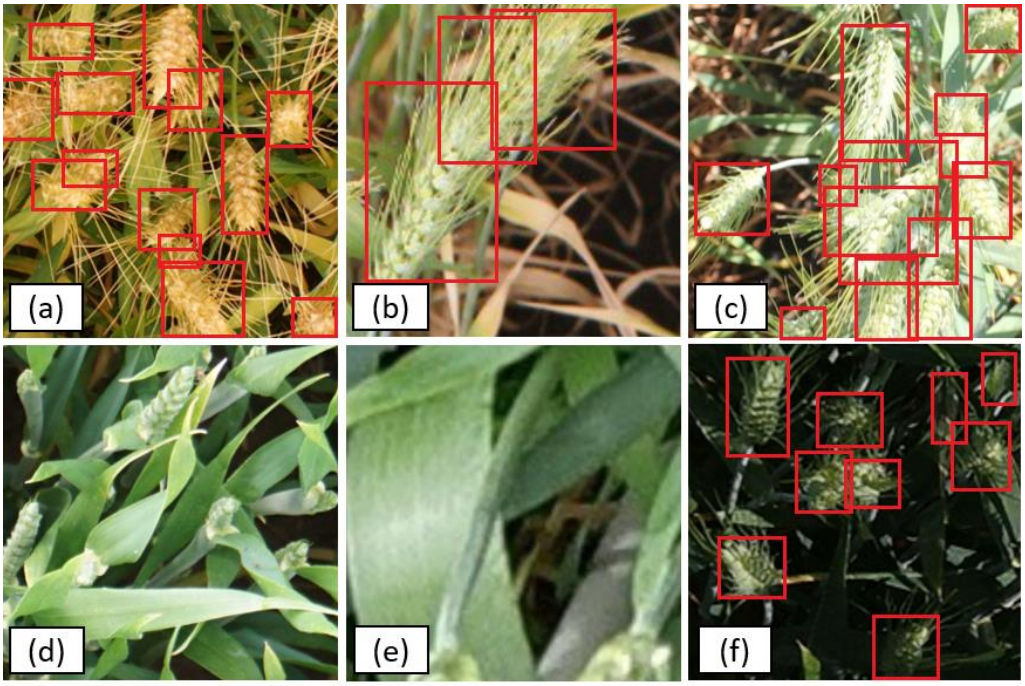

Figure 2: Examples of wheat heads difficult to label. It includes overlapping heads (a-c), heads at emergence (d), heads that are partly cut at border of the image (e), and images with a low illumination (f). These examples are zoomed-in views from images contained in the dataset, with different zoom factors. Note that the image (d) was removed from the dataset because of the ambiguity of heads at emergence.

Several institutions already labelled their sub-datasets. For the datasets not labelled, we used a "weakly supervised deep learning framework" [27] to label images efficiently for these sub-datasets. A YoloV3 model [28] was trained over UTokyo_1 and Arvalis_1 sub-datasets and then applied to the un-labelled sub-datasets. Boxes with an associated confidence score larger than 0.5 were retained and proposed to the user for correction. This semi-automatic active learning increased by a factor of four the throughput of the labelling process as compared to a fully interactive process. The process is detailed in Figure S1.

This first labelling result was then reviewed by two operators independent from the sub-dataset's institution. When too large discrepancies between reviewers were observed, another labelling and reviewing round was initiated. Approximately 20 operators contributed to this labelling effort. This collaborative process and repeated reviews ensure high level of accuracy and consistency across the sub-datasets.

3. Description of the dataset

3.1. General statistics

The GWHD dataset represents 4,698 squared patches extracted from the 2219 original high resolution RGB images acquired across the 11 sub-datasets (Table 3). It represents 188,445 labelled heads which makes on the average 40 heads per image in good agreement with the 20 to 60 targeted heads per image. However, the distribution between and within sub-datasets is relatively broad (Figure 3a). About 100 images contain no heads while few ones contain more than 100 heads with a maximum of 120 heads. Multiple peaks corresponding to the several sub-datasets (Figure 3b) can be observed corresponding mainly to variations in head density that depends on genotypes and environmental conditions. The size of the bounding boxes around the heads shows a slightly skewed gaussian distribution with a median typical dimension of 77 pixels (Figure 3b). The typical dimension is computed as the square root of the area. It corresponds well to the targeted scale, i.e. 1.5 cm x 10 cm approximate head size with an average resolution close to 0.4 mm/pixel which represents a typical dimension of 97 pixels per head, although the simple horizontal area projected does not corresponds exactly to the viewing geometry of the RGB cameras. The harmonization of the scale across sub-datasets can be further confirmed by the visual impression as illustrated in Figure 4.

Sub-dataset name	**Nb. of acquired images**	**Nb. of patch per image**	**Original GSD (mm)**	**Sampling factor**	**Used GSD (mm)**	**Nb. of labelled images**	**Nb. of labelled heads**	**Average nb. of heads/ images**

UTokyo_1	994	1	0.43	1	0.43	994	29174	29
UTokyo_2	30	4	0.6	2	0.3	120	3263	27
Arvalis_1	239	6	0.23	0.5	0.46	1055*	45716	43
Arvalis_2	51	4	0.56	2	0.28	204	4179	20
Arvalis_3	152	4	0.56	2	0.28	608	16665	27
INRAE_1	44	4	0.56	2	0.28	176	3701	21
USask_1	100	2	0.45	1	0.45	200	5737	29
RRes_1	72	6	0.33	1	0.33	432	20236	47
ETHZ_1	375	2	0.55	1	0.55	747*	51489	69
NAU_1	20	1	0.21	1	0.21	20	1250	63
UQ_1	142	1	0.2	0.5	0.4	142	7035	50
total	2219	-	-	-	-	4698	188445	-

Table 3: Statistics for each component of the Global wheat head Detection

*some labeled images have been removed during the labelling process

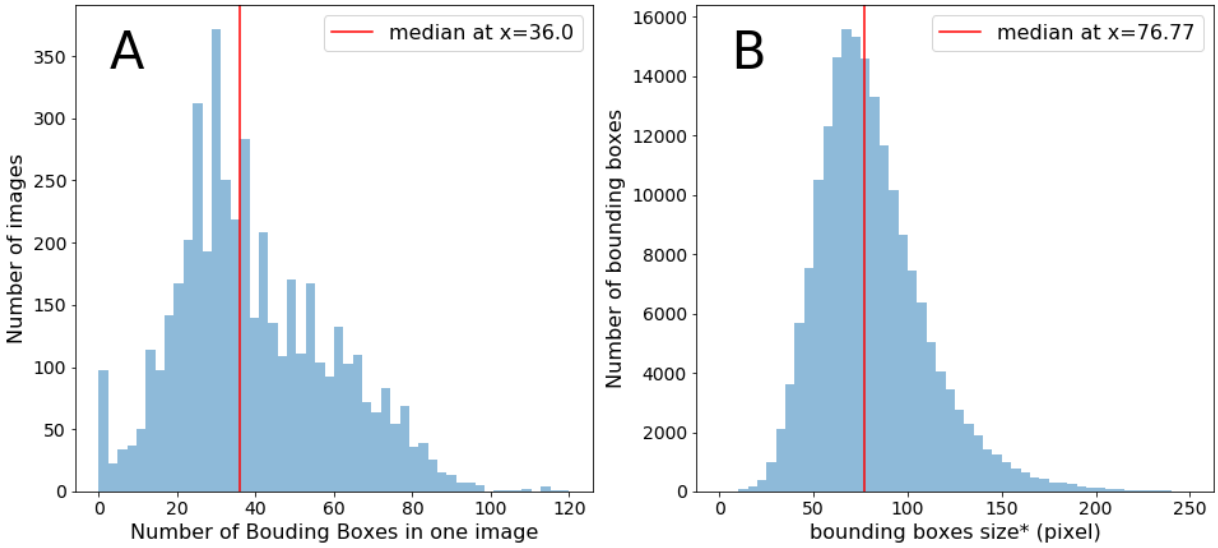

Figure 3: Distribution of the number of bounding boxes per image (a) and bounding boxes size* (b) in the GWHD Dataset.

*the bounding box size is defined as the square root of the bounding box area in pixel.

3.2. Diversity of sampled genotypes, environments, and developmental stages

The diversity of acquisition conditions sampled by the GWHD dataset is well illustrated by Figure 4: illumination conditions are very variable, with a wide range of heads and background appearance. Further, we observe a variability of the ear orientation and view directions, with almost nadir directions up to oblique viewing as in the case of ETHZ_1 (Figure 4). A selection of bounding boxes extracted from the several sub-datasets (Figure 5) shows variation of bounding-box area and aspect ratio, depending on the head orientation and viewing direction. A large diversity of head appearance is observed, with variation in the presence of awns and awn size, head color, as well as blurriness and few heads cut when the bounding box was crossing the edge of the image (Figure 5).

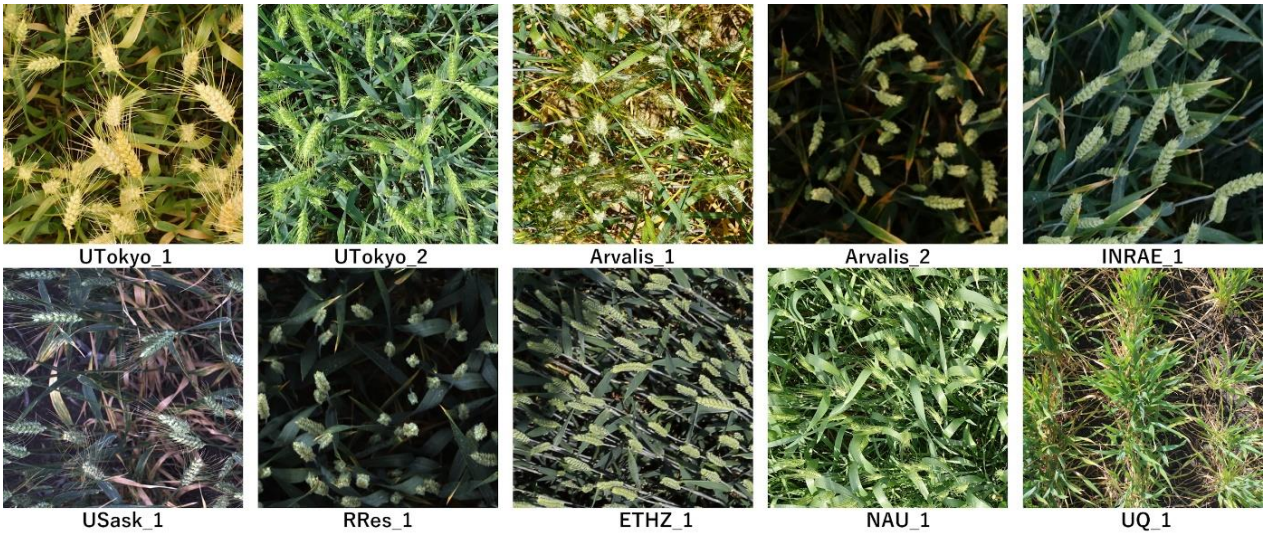

Figure 4: Example images from different acquisition sites after cropping and rescaling.

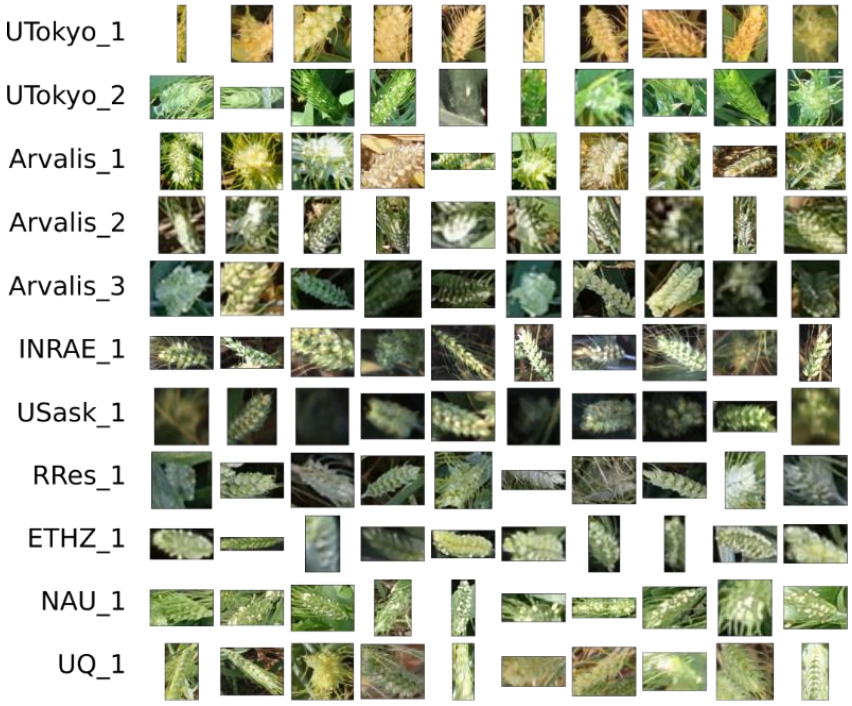

Figure 5: A selection of bounding boxes for each sub-dataset. The same size of pixels is used across all the bounding boxes displayed.

3.3. Comparison to other datasets

Several open-source datasets have already been proposed in the plant phenotyping community. The CVPPP datasets [29] have been widely used for rosette leaf counting and segmentation. The KOMATSUNA dataset also includes segmented rosette leaves, but in time-lapse videos [30]. The Nottingham ACID Wheat dataset includes wheat head images captured in a controlled environment with individual spikelets annotated [31]. However, comparatively few open-source datasets include images from outdoor field contexts, which are critical for the practical application of phenotyping in crop breeding and farming. A few datasets have been published for weed classification [32][33]. The GrassClover Dataset includes images of forage fields and semantic segmentation labels for grass, clover and weed vegetation types [34]. Datasets for counting sorghum [27][35] and wheat heads [36] have also been published with dot annotations.

In terms of phenotyping datasets for object detection, our GWHD dataset is currently the largest open labelled dataset freely available for object detection for field plant phenotyping. MinneApple [37] is the only comparable dataset in terms of diversity in the field of phenotyping but proposes fewer images and less

diversity in terms of location. Other datasets like MS COCO [38] or Open Images V4 [39] are much larger and sample many more object types for a wide range of other applications. The corresponding images usually contain fewer objects, typically less than ten per image (Figure 6). However, some specific datasets like PUCPR [40], CARPK [41], SKU-110K [42] aim at solving the problem of detecting objects (e.g., cars, products) in dense contexts. They have a much higher object density than GWHD Dataset, but with fewer images for PUCPR and CARPK, while SKU-110 contains more images than our GWHD Dataset (Figure 6). The high occurrence of overlapping and occluded objects is unique to the GWHD Dataset. This makes labelling and detection more challenging, especially compared to SKU-110K which does not seem to present any occlusion. Finally, wheat heads are complex objects with a wide variability of appearance as demonstrated previously, surrounded by a very variable background which would constitute a more difficult problem than detecting cars or densely packed products on store shelves.

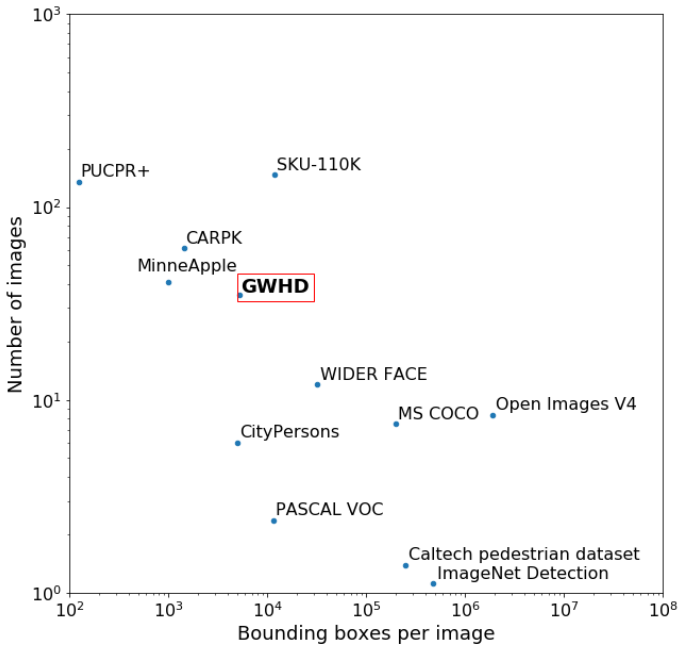

Figure 6: Comparison of GWHD dataset with other object detection datasets. Both axes are in log-scale.

4. Target use case: Wheat Head Detection challenge

The main goal of the dataset is to contribute to solving the challenging problem of wheat head detection from RGB high resolution images. A competition will be held from 4th 2020 to July 2020 to benchmark wheat head detection methods using the GWHD dataset for training and testing (http://www.global-wheat.com/2020-challenge/).

4.1. Split between training and testing datasets

In machine learning studies, it is common to randomly split a dataset into training and testing samples. However, for the GWHD competition, we specifically aim to test the performance of the method for unseen genotypes, environments, and observational conditions. Therefore, we grouped all images from Europe and North America as the training dataset, which covers enough diversity to train a generic wheat head detection model. This training dataset corresponds to 3422 images representing 73% of the whole GWHD dataset images. To evaluate model performance, including robustness against unseen images, the test data set includes all the images from Australia, Japan, and China, representing 1276 images.

4.2. Evaluation Metrics

The choice of bounding boxes as labels allows leveraging the dataset for an object detection dataset. The mean average precision computed from the true and false positives is usually used to quantify detection performances. A true positive corresponds to a predicted bounding box with an intersection over union (IoU) higher than 0.5 with the closer labelled bounding box. A false positive corresponds to a predicted bounding box with IoU<0.5 with the closer labelled bounding box. The mean average precision noted "mAP@0.5" is the metrics considered to evaluate the performances of the several detection methods tested within the wheat head detection challenge organized in 2020. The head detection is mandatory for characterizing the heads in terms of size, inclination, color, or sanitary state. However, the number of wheat heads per image is also a

highly desired trait. Future competitions using the GWHD dataset could focus on wheat head counting with metrics such RMSE, relative RMSE (rRMSE) and R^2 to quantify the performances of a counting methods.

4.3. Baseline method

To set a baseline detection accuracy for the GWHD dataset, we provide results based on a standard object detection method. We trained a two-stage detector, Faster-RCNN, with a ResNet34 as the backbone. Due to memory constraints, the input size was set to 512x512 pixels. For each image in the training dataset, we randomly sampled 10 patches of size 512 x 512 pixels. We predicted on a set of overlapping patches of size 512 x 512 pixels regularly extracted from the test images of size 1024 x 1024 pixels and then merged the results. After 20 epochs, representing 26638 steps the model yielded a mAP@0.5 of 0.68 and a mean RMSE of 13.75 wheat heads per image which corresponds to rRMSE=34%. The relatively poor performance of a standard object detection network on the GWHD dataset, provides an opportunity for substantial future improvement with novel methods.

5. Discussion

5.1. Image acquisition recommendations

To successfully detect wheat heads, they must be clearly visible within the images, with minimum overlap among heads and leaves and fully emerged. For some genotypes and environmental conditions, we observed that wheat heads tend to bend for the latest grain filling stages, increasing the overlap between heads. Conversely, for the stages between heading and flowering, some heads are not yet fully emerged and are therefore difficult to see. It is therefore recommended to acquire images right after flowering when heads have fully emerged, and heads are still upright in the field.

Near nadir viewing directions that limit the overlap between heads are recommended, especially in case of high-density head population. Narrow field of view is therefore preferred. However, narrow field of view may result into small image footprint in case of short distance between the camera and the top of the canopy. It is therefore preferred to increase the camera height to get a larger sampled area while limiting the number of heads that will cross the edge of the image. The size of the sampled area will be critical when head identification is used for estimating the head population density. The minimum sampled area should be that of our squared patch, i.e. 1024x1024 pixels of 0.4 mm/pixel which is about 40 x 40 cm area. To get a maximum field of view of ±15° allowing to minimize head overlapping, the distance between the camera and the top of the canopy should be around 1.0 m. However, larger sampling area should be preferred, especially in the case of head population density estimation, where at least 1.0 m² should be sampled to account for possible heterogeneity across rows. This would be achieved with a 2.5m distance between the camera and the top of the canopy.

When estimating the number of head population density, accurate knowledge of the sampled area is critical. The non-parallel geometry of image acquisition, with significative panoramic effects induces uncertainties on the sampled area: even for our typical case with limited panoramic effect (±15° field of view), for an image acquired at 2.5. m from the top of the canopy, an error of 10 cm in canopy height estimation induces 8% error in the sampled area which directly transfers to the head density estimation. Further, the definition of the reference level to compute the sampled area is still an open question since the head layer is generally thicker than 25 cm, which induces a 21% difference in the sampled area between the top and the bottom head layer. Further work should investigate this important question.

Finally, our experience suggests that using a sub-millimetric resolution at the top of the canopy is required for efficient head detection. However, the optimal resolution is still to be defined. Previous work [8] was recommending 0.3 mm GSD, while the GWHD dataset includes GSD ranging from 0.28 to 0.55 mm. Further work should investigate this important aspect, particularly regarding the possibility to use UAV observations for head density estimation in large phenotyping platforms.

5.2. Minimum information associated to the sub-datasets and FAIR principles

The FAIR principles (Findable, Accessible, Interoperable and Reusable [43] should be applied to the images that populate the GWHD dataset. To verify the FAIR principles, a minimum set of metadata should be associated to each image as proposed in [44]. The lack of metadata was an issue for precise data harmonization and is limiting for further interpretation [45] and possible meta-analysis. We therefore recommend attaching a minimum set of information to each image and sub-dataset. In our case, a sub-dataset generally corresponds to an image acquisition session, i.e. a series of images acquired over the same experiment on the same date and with the same camera. the experiment metadata are all the metadata related to agronomic characteristics

of the session, the acquisition metadata are all the metadata related to the vector and the sensor. Both can be defined at the session level and the image level. Our recommendations are summarized in Table 4.

	Session level	Image level
Experiment metadata	Name of the experiment (PUID) † Name of institution GPS coordinates (°) Email address of the contact person Date of the session (yyyymmdd) Wheat species (durum, aestivum …)* Development stage / ripening stage*	Microplot id Row spacing Sowing density Name of the genotype (or any identifier) † presence or not of awns.
Acquisition metadata	**Vector characteristics**: Name Type (handheld, cart, phenomobile, gantry, UAV) Sampling procedure Distance to the ground (m)* **Camera characteristics,** Model, Focal length of lens (mm), Size of the pixel at the sensor matrix (µm) Sensor dimensions (pixels x pixels), **Camera settings**, Mode (automatic, speed or aperture priority, manual) White balance correction	Camera aperture Shutter speed ISO Canopy height (m) Position of the image in the microplot§

Table 4: The minimum metadata that should be associated to images of heads

* this may be alternatively reported at the image level if it is variable within a session

† persistent unique identifier (PUID). This may be a DOI as for genetic resources regulated under the on Plant Genetic Resources for Food and Agriculture (https://ssl.fao.org/glis) or any other identifier including the information of the maintainer of the genetic material, ripening stage

§ In case of multiple images over the same microplot.

5.3. Need for GWHD expansion

The innovative and unique aspect of the GWHD dataset is the significant number of contributors from around the world, resulting in a large diversity across images. However the diversity within each continent and environmental conditions is not well covered today: more than 68% of the GWHD dataset come from Europe and 43% from France. Further, some regions are currently missing, including Africa, Latin America, the Middle East. An expansion of the GWHD dataset is therefore highly desirable to get a more comprehensive dataset. We therefore invite potential contributors to complement the GWHD dataset with their sub-datasets. The proposed guidelines for image acquisition and the associated metadata should be followed to keep a high level of consistency and respect the FAIR principles. We therefore encourage potential contributors to contact the corresponding authors through www.global-wheat.com

6. Conclusion

Object detection methods for identifying wheat heads in images is useful for head population density estimation. However, we showed that an accurate knowledge of the reference level to compute the image footprint area was mandatory. More work is needed to solve this important question. However, head detection may be also considered as a first step in the search for additional traits characterizing the heads such as the spatial distribution between rows, the presence of awns, size, inclination, color, grain filling stage, or sanitary state. These traits may be useful for the breeders, but some of these traits including the sanitary state may also serve the needs of farmers to better manage their crops.

The Global Wheat Head Detection Dataset is an extensive and diverse dataset for wheat head detection and localization. It is designed to develop, and benchmark head detection methods proposed by the community. It represents a large collaborative international effort. An important aspect gained through the compilation of diverse sub-datasets was to propose guidelines for image acquisition, minimum metadata to respect the FAIR principles, and guidelines and tools for labelling the heads. These guidelines will help practitioners to get more consistent sub-datasets that could ultimately contribute to expand the GWHD dataset. The GWHD dataset collaborative effort is expected to represent a major contribution for advancing high throughput detection of wheat heads by letting many teams to compete for finding the more accurate and robust method. The solutions proposed in the competition will be made open-source and shared with the plant phenotyping community to give access to a larger number to state-of-the-art methods.

Acknowledgments

The French team received support from ANRT for the CIFRE grant of Etienne David, co-funded by Arvalis. The study was partly supported by several projects including ANR PHENOME, ANR BREEDWHEAT, CASDAR LITERAL and FSOV "Plastix". Many thanks to the people who annotated the French datasets, including Frederic Venault, Xiuliang Jin, Mario Serouard, Ilias Sarbout, Carole Gigot, Eloïse Issert, Elise Lepage.

The Japanese team received support from JST CREST Grant Number JPMJCR16O3, JPMJCR16O2, JPMJCR1512 and MAFF Smart-breeding system for Innovative Agriculture (BAC1003), Japan. Many thanks to the people who annotated the Japanese dataset, including Kozue Wada, Masanori Ishii, Ryuuichi kanzaki, Sayoko Ishibashi, Sumiko Kaneko.

The Canadian team received funding from the Plant Phenotyping and Imaging Research Center through a grant from the Canada First Research Excellence Fund. Many thanks to Steve Shirtliffe, Scott Noble, Tyrone Keep, Keith Halco, and Craig Gavelin for managing the field site and collecting images.

Rothamsted Research received support from the Biotechnology and Biological Sciences Research Council (BBSRC) of the United Kingdom as part of the Designing Future Wheat (BB/P016855/1) project. We are also thankful to Prof Malcolm J Hawkesford who leads the DFW project and Dr Nicolas Virlet for conducting the experiment at Rothamsted Research.

The Gatton, Australia dataset was collected on a field trial conducted by CSIRO and UQ, with trial conduct and measurements partly funded by the Grains Research and Development Corporation (GRDC) in project CSP00179. A new GRDC project involves several of the authors and supports their contribution to this paper.

The dataset collected in China was supported by the Program for High-Level Talents Introduction of Nanjing Agricultural University (440—804005). Many thanks to Jie Zhou and many volunteers from Nanjing Agricultural University to accomplish the annotation.

The data set collection at ETHZ was supported by Prof. Achim Walter who leads the Crop Science group. Many thanks to Kevin Keller for the initial preparation of the ETHZ dataset, and Lara Wyser, Ramon Winterberg, Damian Käch, Marius Hodel and Mario Serouard (INRAE) for annotation of the ETHZ dataset and to Brigita Herzog and Hansueli Zellweger for crop husbandry.

Author contributions:

E.D., S.M., B.S, F.B. organized field experiment and data collection for France dataset. P.S.T organized field experiment and data collection for U.K. dataset. H.A., N.K, A.H. organized field experiment and data collection for Switzerland dataset. G.I., K.N., W.G. organized field experiment and data collection for Japan dataset. S.L., F.B. organized field experiment and data collection for China dataset. C.P., M.B., I.S. organized field experiment and data collection for Canada dataset. B.Z., S.C.C organized field experiment and data collection for Australia dataset. E.D and S.M harmonized the sub-datasets. W.G, E.D. and S.M. built the initial Wheat Head Detection model and conducted pre-labelling process. E.D. administered the labelling platform and all authors contributed to data labelling and quality check. E.D. built the baseline model for the competition. E.D. and S.M. wrote the first draft of the manuscript, they contributed equally to this work. All authors gave input and approved the final version.

Competing interests: The authors declare that there is no conflict of interest regarding the publication of this article.

Data Availability: The GWHD Dataset will be public available online at (www.global-wheat.com) under MIT licenses.

Supplementary Materials

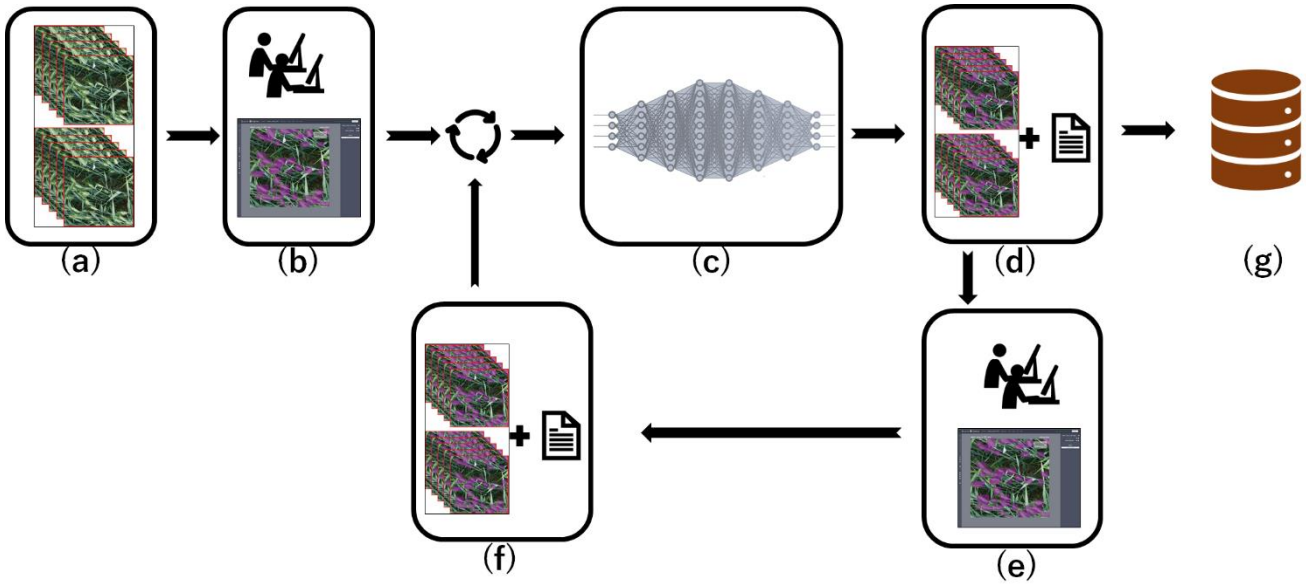

Figure S1: the proposed "weakly supervised deep learning framework" to pre-label images efficiently.

(a) Original image input.
(b) Labelling work for randomly selected original images.
(c) Train an initial model with labeled image from (b).
(d) Apply the initial model on original images left.
(e) Feed the generated labels and the image into the image annotator app for validating the bounding box locations and corrections by a human annotator.
(f) Add corrected labels and the image to data pool (b)
(g) Acquire the final annotated dataset after iterating the process (a)~(f)~(d) until reached to desired performance as determined by evaluating the trained model at the end of every iteration.

References

[1] M. P. Reynolds and N. E. Borlaug, "Applying innovations and new technologies for international collaborative wheat improvement," Journal of Agricultural Science, vol. 144, no. 2. Cambridge University Press, pp. 95–110, Apr-2006.

[2] N. Brisson, P. Gate, D. Gouache, G. Charmet, F. X. Oury, and F. Huard, "Why are wheat yields stagnating in Europe? A comprehensive data analysis for France," F. Crop. Res., vol. 119, no. 1, pp. 201–212, Oct. 2010.

[3] B. Schauberger, T. Ben-Ari, D. Makowski, T. Kato, H. Kato, and P. Ciais, "Yield trends, variability and stagnation analysis of major crops in France over more than a century," Sci. Rep., vol. 8, no. 1, pp. 1–12, Dec. 2018.

[4] M. Reynolds et al., "Breeder friendly phenotyping," Plant Science. Elsevier Ireland Ltd, p. 110396, Jan-2020.

[5] J. Crain, S. Mondal, J. Rutkoski, R. P. Singh, and J. Poland, "Combining High‐Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding," Plant Genome, vol. 11, no. 1, pp. 1‐14, Mar. 2018.

[6] A. Hund, L. Kronenberg, J. Anderegg, K. Yu, A. W.-A. in crop Breeding, and U. 2019, Non-invasive phenotyping of cereal growth and development characteristics in the field. Cambridge, 2019.

[7] A. Walter, F. Liebisch, and A. Hund, "Plant phenotyping: from bean weighing to image analysis," Plant Methods, vol. 11, no. 1, p. 14, Mar. 2015.

[8] S. Madec, X. Jin, H. Lu, B. De Solan, S. Liu, F. Duyme, E. Heritier, and F. Baret, "Ear density estimation from high resolution RGB imagery using deep learning technique," Agric. For. Meteorol., vol. 264, pp. 225–234, Jan. 2019.

[9] M.M Hasan., J.P. Chopin, H. Laga, and S.J Miklavcic, "Detection and analysis of wheat spikes using convolutional neural networks", Plant Methods, 14(1), 100, 2018

[10] M.Z. Alom, T.M Taha, C. Yakopcic, S. Westberg,P. Sidike, M. S. Nasrin, M. Hasan, B.C Van Essen, A. Awwal and V.K Asari, "A state-of-the-art survey on deep learning theory and architectures," Electron., vol. 8, no. 3, 2019.
[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Berstein, A. C. Berg and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015.
[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 386–397, Feb. 2020.
[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015, vol. 9351, pp. 234–241.
[15] S. Aich and I. Stavness, "Global Sum Pooling: A Generalization Trick for Object Counting with Small Datasets of Large Images."
[16] H. Xiong, H. Lu, C. Liu, L. Liu, Z. Cao, and C. Shen, "From Open Set to Closed Set: Counting Objects by Spatial Divide-and-Conquer," Aug. 2019.
[17] M. P. Pound, J. A. Atkinson, D. M. Wells, T. P. Pridmore, and A. P. French, "Deep Learning for Multi-task Plant Phenotyping," bioRxiv, p. 204552, Oct. 2017.
[18] J. R. Ubbens and I. Stavness, "Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks," Front. Plant Sci., vol. 8, p. 1190, Jul. 2017.
[19] A. K. Singh, B. Ganapathysubramanian, S. Sarkar, and A. Singh, "Deep Learning for Plant Stress Phenotyping: Trends and Future Perspectives," Trends Plant Sci., vol. xx, pp. 1–16, 2018.
[20] P. Sadeghi-Tehran, N. Virlet, E. M. Ampe, P. Reyns, and M. J. Hawkesford, "DeepCount: In-Field Automatic Quantification of Wheat Spikes Using Simple Linear Iterative Clustering and Deep Convolutional Neural Networks," Front. Plant Sci., vol. 10, no. September, pp. 1–16, 2019
[21] J. A. Gibbs, A. J. Burgess, M. P. Pound, T. P. Pridmore, and E. H. Murchie, "Recovering Wind-Induced Plant Motion in Dense Field Environments via Deep Learning and Multiple Object Tracking," Plant Physiol., vol. 181, no. 1, pp. 28–42, Sep. 2019.
[22] A. Milioto, P. Lottes, and C. Stachniss, "Real-time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs," 2017.
[23] J. Ubbens, M. Cieslak, P. Prusinkiewicz, and I. Stavness, "The use of plant models in deep learning: An application to leaf counting in rosette plants," Plant Methods, vol. 14, no. 1, Jan. 2018.
[24] J. Anderegg, K. Yu, H. Aasen, A. Walter, F. Liebisch, and A. Hund, "Spectral Vegetation Indices to Track Senescence Dynamics in Diverse Wheat Germplasm," Front. Plant Sci., vol. 10, p. 1749, Jan. 2020.
[25] X. Jin, S. Madec, D. Dutartre, B. de Solan, A. Comar, and F. Baret, "High-Throughput Measurements of Stem Characteristics to Estimate Ear Density and Above-Ground Biomass," Plant Phenomics, vol. 2019, p. 4820305, May 2019.
[26] J. Brooks, "COCO Annotator." 2019.
[27] S. Ghosal, B. Zheng, S. C. Chapman, A.B. Potgieter, D.R. Jordan, X. Wang, A.K. Singh, A. Singh, M. Hirafuji, S. Ninomiya, B. Ganapathysubramanian, S. Sarkar and W. Guo, "A Weakly Supervised Deep Learning Framework for Sorghum Head Detection and Counting," Plant Phenomics, vol. 2019, pp. 1–14, 2019.
[28] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," Apr. 2018.
[29] H. Scharr, M. Minervini, A. Fischbach, and S. A. Tsaftaris, "Annotated Image Datasets of Rosette Plants," pp. 1–16, 2014.
[30] Hideaki Uchiyama, Shunsuke Sakurai, Masashi Mishima, Daisaku Arita, Takashi Okayasu, Atsushi Shimada and Rin-ichiro Taniguchi, "An easy-to-setup 3D phenotyping platform for KOMATSUNA dataset," ICCV Workshop on Computer Vision Problems in Plant Phenotyping, pp.2038-2045, 2017.
[31] M.P. Pound, J.A Atkinson, D.M. Wells, T.P. Pridmore & A.P French, "Deep learning for multi-task plant phenotyping.", In Proceedings of the IEEE International Conference on Computer Vision Workshops", pp. 2055-2063, 2017
[32] I. Sa, M. Popović, R. Khanna, Z. Chen, P. Lottes, F. Liebisch, J. Nieto,, C. Stachniss, A. Walter and R. Siegwart, R., "Weedmap: a large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming", Remote Sensing, 10(9), 1423,2018
[33] N. Teimouri, M. Dyrmann, P.R. Nielsen, S.K. Mathiassen, G.J. Somerville and R.N. Jørgensen, "Weed growth stage estimator using deep convolutional neural networks", Sensors, 18(5), 1580, 2018
[34] S. Skovsen, M. Dyrmann, A.K. Mortensen, M.S. Laursen, R. Gislum, J. Eriksen, S. Farkhani, H. Karstoft and R.N.Jorgensen, "The GrassClover Image Dataset for Semantic and Hierarchical Species Understanding in Agriculture", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019

[35] W. Guo, B. Zheng, A.B Potgieter, J. Diot, K. Watanabe, K. Noshita, D.R. Jordan, X. Wang, J. Watson, S. Ninomaya and S.C. Chapman, “Aerial imagery analysis–quantifying appearance and number of sorghum heads for applications in breeding and agronomy”, Frontiers in plant science, 9, 1544, 2018.
[36] H. Xiong, Z. Cao, H. Lu, S. Madec, L. Liu, and C. Shen, “TasselNetv2: In-field counting of wheat spikes with context-augmented local regression networks,” Plant Methods, vol. 15, no. 1, 2019.
[37] N. Hani, P. Roy, and V. Isler, “MinneApple: A Benchmark Dataset for Apple Detection and Segmentation,” IEEE Robot. Autom. Lett., vol. 5, no. 2, pp. 852–858, Apr. 2020.
[38] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, “Microsoft COCO: Common objects in context,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8693 LNCS, no. PART 5, pp. 740–755,2014
[39] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duering and V. Ferrari, “The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale,” Int. J. Comput. Vis., Nov. 2018.
[40] P. R. L. De Almeida, L. S. Oliveira, A. S. Britto, E. J. Silva, and A. L. Koerich, “PKLot-A robust dataset for parking lot classification,” Expert Syst. Appl., vol. 42, no. 11, pp. 4937–4949, Jul. 2015.
[41] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, “Drone-based Object Counting by Spatially Regularized Regional Proposal Network,” Proc. IEEE Int. Conf. Comput. Vis., vol. 2017-October, pp. 4165–4173, Jul. 2017.
[42] E. Goldman, R. Herzig, A. Eisenschtat, J. Goldberger, and T. Hassner, “Precise detection in densely packed scenes,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2019-June, pp. 5222–5231, 2019
[43] C. Pommier, C. Michotey, G. Cornut, P. Roumet, E. Duchêne, R. Flores, A. Lebreton, M. Alaux, S. Durand and E. Kimmel "Applying FAIR principles to plant phenotypic data management in GnpIS." Plant Phenomics 2019: 1671403, 2019
[44] H. Ćwiek-Kupczyńska, T. Altmann, D. Arend, E. Arnaud, D. Chen, G. Cornut, F. Fiorani, W. Frohmberg, A. Junker and C. Klukas, "Measures for interoperability of phenotypic data: minimum information requirements and formatting." Plant Methods 12(1): 44, 2016
[45] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, “Subcategory-Aware convolutional neural networks for object proposals & detection,” in Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, pp. 924–933, 2017

.