The challenge of robust trait estimates with Deep Learning on high resolution RGB images

Etienne David

AVIGNON UNIVERSITÉ

The challenge of robust trait estimates with Deep Learning on high resolution RGB images

Le défi de l'estimation robuste des traits avec l'apprentissage profond sur des images RVB à haute résolution

Etienne David[1,2]

Thèse soutenue le 2 novembre 2021

[1] Institut national de recherche pour l'agriculture, l'alimentation et l'environnement - UMR EMMAH 114
[2] Arvalis - Institut du Végétal

ED 536 - Spécialité Sciences Agronomiques

Composition du jury

Pr Tony Pridmore (University of Nottingham)	Rapporteur
Dr Alexis Joly (LIRMM, INRIA)	Rapporteur
Dr Karine Chenu (University of Queensland)	Examinatrice
Dr Marion Prudent (UMR Agroécologie - INRAe)	Examinatrice
Dr Frédéric Baret (UMR EMMAH, INRAE)	Directeur
Benoit de Solan (Arvalis)	Encadrant

INRAE CAPTE ARVALIS Institut du végétal

Abstract

Le défi des estimations robustes de trait par Deep Learning à partird'imagerie haute résolution RVB

Le phénotypage à haut débit des plantes, notamment dans le cadre d'acquisitions en plein champ, repose sur l'interprétation de données issues de différents capteurs mis en œuvre sur des vecteurs variés tels que des tracteurs, des robots ou des drones. Initialement, ces données étaient interprétées à l'aide d'algorithmes de télédétection exploitant la résolution spectrale du signal. Mais depuis 2015, les progrès du "Deep Learning", basé sur l'entrainement à partir d'exemples, ont permis des résultats prometteurs pour mesurer des traits essentiels comme le taux de couverture ou le comptage de plantes ou d'organes. Ces algorithmes utilisent des couches de convolution apprises, permettant de tirer parti de l'organisation spatiale du signal. L'avantage de ces méthodes est qu'elles sont basées sur des capteurs Rouge-Vert-Bleu (RVB), qui sont beaucoup moins coûteux que les imageurs multi- ou hyperspectraux. Cependant, les algorithmes de Deep Learning sont sensibles aux changements de la distribution entre les données utilisées pour l'entrainement et les données prédites. En pratique, des erreurs de prédiction variables et non prédictibles d'un site à l'autre peuvent être observées. L'objectif de la thèse est de comprendre les causes de ces variations et de proposer des solutions pour des estimations de traits phénotypiques fiables en utilisant le Deep Learning. L'étude porte sur la détection de plantes et d'organes à partir d'images RVB haute résolution acquises sur le terrain. Nos travaux ont d'abord porté sur la constitution de bases de données d'images diversifiées provenant de différents lieux et stades de développement pour l'émergence de plantes (maïs, betterave, tournesol) et les épis de blé, ce qui a permis la publication de deux bases de données annotées, regroupant 27 sessions d'acquisition pour le drone et 47 pour la détection d'épis. Ces jeux de données démontrent la différence de performances entre les résultats publiés et les nôtres en raison du changement de distribution. Pour dépasser les limites des méthodes habituelles, nous avons organisé deux concours de données, les Global Wheat Challenges, en 2020 et 2021, qui nous ont permis d'obtenir des solutions entraînées pour la robustesse sur un jeu de données différent de celui de l'entraînement. L'analyse des solutions a montré l'importance des stratégies d'entraînement pour la robustesse au-delà des architectures utilisées. Nous avons également montré que ces solutions peuvent être déployées efficacement en remplacement du comptage manuel. Enfin, nous avons démontré l'inefficacité des fonctions d'entraînement conçues pour l'entraînement robuste. Notre travail ouvre la perspective d'une meilleure évaluation du Deep Learning dans le contexte du phénotypage à haut débit et donc de la confiance dans son utilisation en conditions réelles.

The challenge of robust trait estimates with DeepLearning on high resolution RGB images

High throughput plant phenotyping, especially in the context of open field acquisitions, relies on the interpretation of data from different sensors implemented on various vectors such as tractors, robots or drones. Initially, these data were interpreted using remote sensing algorithms that exploit the spatial resolution of the signal. Since 2015, however, progresses of "Deep Learning", based on the training on examples, has already obtained promising results for measuring the rate of cover, counting plants or organs. It uses learned convolution layers, can take advantage of the spatial organization of the signal. The advantage of these methods is that they are based on Red-Green-Blue (RGB) sensors, which are much less expensive than multi- or hyperspectral imagers. However, these methods are sensitive to changes in the distribution between the data used in training and the predicted data. In practice, variable prediction errors from site to site can be observed using these methods.

The objective of the thesis is to understand the causes of these variations and propose solutions for reliable phenotypic trait estimates using Deep Learning. The study focuses on detecting plants and organs from high-resolution RGB images acquired in the field. Our work first focused on the constitution of diversified image databases from different locations and stages of development for plant emergence (maize, beet, sunflower) and wheat ears, which allowed the publication of two annotated databases, grouping 27 acquisition sessions for the drone and 47 for the ear detection. The datasets demonstrate the performances difference between the published results and ours due to the change in distribution. To go beyond the limits of the usual methods, we organized two data competitions, the Global Wheat Challenges, in 2020 and 2021, which allowed us to obtain solutions trained for robustness on a different data set than the training one. The analysis of the solutions showed the importance of the training strategies for robustness beyond the architectures used. We have also shown that these solutions can be effectively deployed as a replacement for manual counting. Finally, we have demonstrated the inefficiency of training functions designed for robust training. Our work opens the prospect of a better evaluation of Deep Learning in the context of high-throughput phenotyping and thus of confidence in its use in real-life conditions.

À V.D. et C.D.

Acknowledgements

A PhD is a transformative journey that was completely different from what I expected. Far from being a solitary experience, it was, in my case, the result of a collaborative effort.

My first thanks go to Fred, my thesis director, and Benoit, my supervisor at Arvalis. Fred, working with you was sometimes far from easy, but fun. I learned from you to be a creative and curious researcher but also rigorous, demanding and hardworking. Always with a smile and good humour, of course! The freedom of research I have had with you is a great gift. Benoit, you have been much more than a mentor. Your sincere benevolence helped me to overcome the moments of doubt during the thesis, and, thanks to your teaching, wheat and maize no longer hold any secrets for me!

Wei, Scott and Ian, my thesis project would never have had as much impact without our collaboration around the Global Wheat Head Dataset and challenge. November 2019 was the crucial turning point between IPPS Nanjing and my visit to Tokyo just before the COVID. Your support has enabled me to go beyond the walls of my laboratory and I owe you a lot for that. This manuscript only exists thanks to the work of all the many co-authors: thanks again to all!

Alexis, Marie, Raùl, Sylvain and Jérémy, I sincerely believe that our discussions helped me become a better researcher within the UMT CAPTE, the perfect place to work on research and industrial issues. Finally, my success of my CIFRE is due to Arvalis and the CIGALE team, which allows me to focus on research while providing all the needed data. Gaëtan, Samuel, Simon, Mario, Eloise, Elise, Lucas, Guillaume, Guy, thanks for sharing this adventure with me!

Contents

List of Figures

List of Tables

1 Introduction

1.1 Agriculture must adapt shortly to face the future challenges

1.1.1 Climatic, environmental, and anthropic constraints on the food system.

Our food system results from the interactions between food production and the consumers [1]. It is challenged by its strong impact on the environment, including global warming: some of the hottest years (2018, 2019,2020) recorded in the world (6th, 3rd and 1st) [2] and in France (2nd, 4th and 1st) [3], happened during the writing of this thesis. Agriculture can also mitigate climate change by sequestering carbon in the soil [4]. Despite the numerous political summits, there is no indication that the trends will reverse in the future, and the average yield of major crops will reduce by 3% to 6.0% for a one degree Celsius increase [5], [6]. The expected increase of +4°C for 2100 is associated with a decrease of 22% for wheat, 12.1% for rice, 26.4% for maize and 1.8% for soybean.

Further, extreme climatic events that can destroy all crops at a country scale are expected to become more frequent. Our food system faces additional challenges: the global population is expected to jump from 7,8 billion people in 2020 to 11 billion in 2100. It requires an increase of 40% of the land area to be cultivated at constant productivity, although cropland represents already 38% of the land surface according to FAO [7]. The diet quality needs to evolve: 39% of adults aged over 18 years old were considered as overweighed in 2016 according to WHO [8]. This phenomenon is described as a "nutrition transition" [9] resulting from the larger quantity of food accessible to the population with more carbohydrates, fats and oils, and combined with a lack of activity. The quality of the diet needs to be addressed for all in the future. At the same time, hunger is still a main issue [10] with still 690 million humans (8.9% of the global population) suffering from too limited access to food in 2016 [11].

At the same time, agriculture should reduce its environmental footprint with less use of pesticides, herbicides or artificial fertilizers, not only in developed countries but also developing ones [12]. Current industrial agriculture practices harm biodiversity while agriculture can adopt new practices emphasizing the role of biodiversity [13]. The extensive use of chemical inputs impacts also human health [14]. Further, the use of genetic modified crop (GMO) [15][16][17][18] is still questioned regarding possible negative impacts on human health and the environment. GMOs are largely used in North America, while most of them are regulated in the European Union.

To summarize, in the next decades, the global food system is required to produce more or to keep the current productivity, with better and more nutritional diets, sequestrate CO2 with a changing climate, less agricultural inputs.

1.1.2 Productivity, potential yield and genetic gain

Food production increases during the 20th century were achieved mainly with a massive rise in crop productivity while the cropland area remained mostly stable [19]. Therefore, the production of food is primarily controlled by crop productivity, i.e. mass of harvested organs per unit area, most often called yield. However, the harvested organs must have sufficient quality to be transformed into food products: wheat requires a minimum protein content to make bread and pasta. Yield results from the accumulation of plant carbohydrates in the harvested organs generally corresponding to the reproductive organs or the root. Photosynthesis processes capture the incoming light to convert water and CO2 into O2 released in the atmosphere and carbohydrates accumulated in the plant (Figure 1). Evans defined the potential yield as "the yield of a cultivar when grown in environments to which it is adapted, with nutrients and water non‐limiting and with pests, diseases, weeds, lodging, and other stresses effectively controlled" [20]. It corresponds to the maximum yield given a

quantity of light, temperature for a shared genotype genetic.

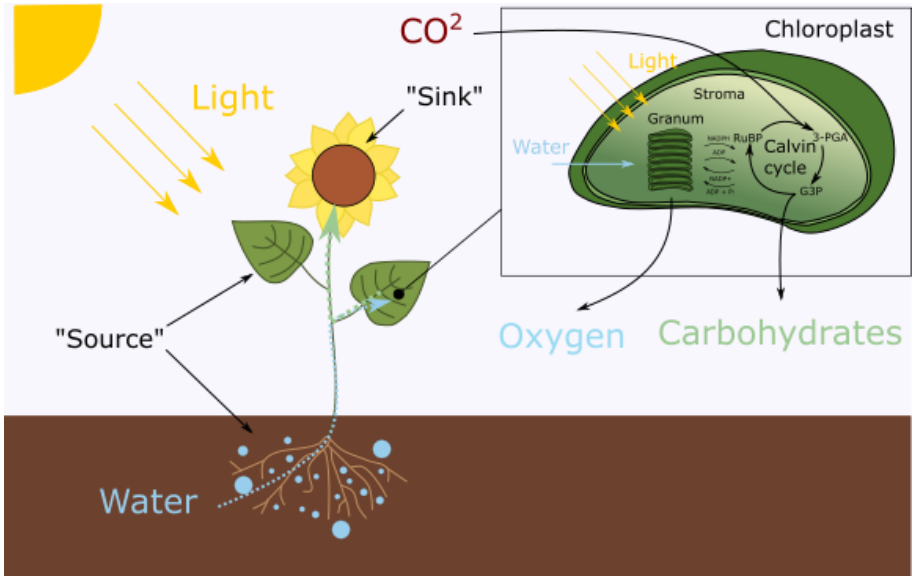


Figure 1. Schematic presentation of the photosynthesis. Chloroplast are cells contains in the leaves that are able to convert CO2 and water to 02 and carbohydrates, that are accumulated in reproductive organs

To keep the yield close to its potential requires avoiding any nutrients and water limitations. Nutrients such as nitrogen, phosphate, and potassium are essential to plant protein production, including those involved in photosynthesis. They can be found in the soil and supplemented by fertilization. Water is the primary source of protons for the Calvin cycle that converts photons into chemical energy (ADP). Water is also necessary to transport nutrients from the soil to the leaves. Diseases (bacterial, virus or fungi) may induce substantial yield loss when they develop in the field and often require using fungicides or bactericides when no natural alternatives can control diseases. Keeping crop yield high requires limiting the competition for light, water and nutrients from the weeds and limiting damages from insects and animals or extreme climatic events such as hail, heatwave or frost. Innovative farming strategies attempt to address all these issues to keep the yield close to its potential. However, the potential yield can also be increased by improving plant functions: a better efficiency of the photosynthesis, better root efficiency in capturing nutrients in the soil, a better leaf orientation to capture light or limit transpiration, high resistance to pests, an increased fraction (harvest index) of the plant dry mass into the harvested organs. For a given environment, most of these functions are governed by the plant genome. Gain of yield potentials with genetic improvement is called genetic gain.

1.1.3 Crop domestication and Scientific revolutions: a brief history of yield improvement science

Humans improved crop yield through several steps driven by key scientific breakthroughs. The genome of modern crops results from these processes, summarized in Figure 2. The first step was the crop domestication[21], which started at least 10.000 years ago for crops such as wheat [22] , maize [23] , rice [24], [25], or sorghum [26][27][28]. This process happened independently at different places such as northeastern America, Mesoamerica, central mid-altitude Andes, West African sub-Sahara, east Sudanic Africa, Near East, northern China, Yangtze China. Domestication is a long process that consists of random crosses between races or close species. Domestication of wilds species is a spectacular example of plant-animal co-evolution: it is even considered that wheat has domesticated humans

[29]. Surprisingly, this process usually requires few chromosomic changes despite spectacular changes in appearance and performance.

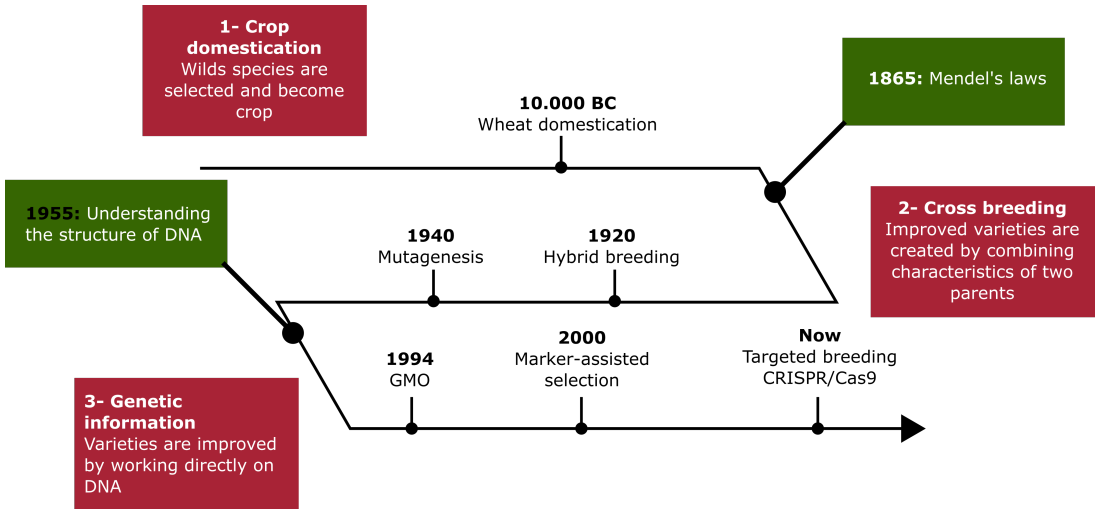


Figure 2. Milestones in crop breeding

The second stage started with the discovery and understanding of Mendel's Laws of genetic. The breakthrough has paved the way for more structured breeding that took into account the parents' characteristics and led to the development of a more formal science of plant breeding which started to be popular during the end of the 19th century, using field trials to select the best performing crosses. The first company, Gartons Agricultural Plant Breeders, was the first to commercialize a new variety of plants thanks to cross-pollination.

From the start of scientific breeding to the end of the second world war, several innovations emerged: cross-breeding (crossing two different varieties), hybrid breeding, or mutation breeding. The development of statistics to handle the effect of the environment during trials was also a critical innovation, which gave birth to modern statistics with the work of Ronald Fisher, in charge of the analysis of the Rothamsted Station [30], near London. Varieties with better yield, disease resistance, and easier harvesting were created thanks to these scientific revolutions. Cross-breeding helped to introduce genes from one cultivar to another. Mutation breeding provokes mutation in plant cells to create new, possibly unusual characters. During this period, the harvest index was improved by reducing the height of stems thanks to cross-breeding. It allows more biomass to go in the grain, less lodging, and easier harvesting. The differences between wheat anterior to this period that is still taller and modern wheat cultivars are presented in Figure 3.

The "Green Revolution", which started in Mexico in the 60s and then spread to other places such as India and Pakistan, was based on creating new varieties of wheat, the most consumed crop, thanks to crop breeding. It introduces genetic gain with new cultivars based on a semi-dwarf gene and a rust-resistant gene. Rust was a significant disease that induced hunger on a large scale. The Green Revolution also introduced more dense sowing density combined with intensive use of fertilizer. Similar approaches were introduced for rice [31] and maize [32]. Despite being already a combination of several innovations, the Green Revolution faces limitations for the identified challenges that agriculture will face in the coming decades. The next revolution would require better exploitation of the ecological services to reduce the environmental impact and create resilient farming systems while producing healthier food.

1.1.4 A change in crop management and genetic improvement are required

With current technologies, reducing negative externalities, making our system more resilient and producing enough food is probably not possible. Some crops rely on the use of few key

Figure 3. Wheat varieties, 19th century to present, John Innes Centre

inputs that, if removed, call into question the system's economic sustainability. The recent sugar beet crisis in France, which took place in 2020, is a good example. Sugar beet is one of the most productive plants with a potential of 24T of sugar per hectare [33], to compare to bread wheat which has only a potential between 8-12T per hectare. It has been grown in France since the 19th century as a replacement for sugar cane. This crop is sensitive to the Beet Chlorosis Virus and Beet Mild Yellowing Virus, that are transmitted by insect and cause a loss of chlorophyll, an essential pigment for the photosynthesis.

Neonicotinoids allowed to grow sugar beet without the viruses' presence during the last two decades, preventing spraying pesticides during the growing season. The pesticides coat the sugar beet seed and spread during the season. In comparison to former solutions, neonicotinoids seemed more reliable and less harmful to the environment. However, the molecule is highly remanent and accumulates in the soil, affecting other micro-organisms that the target pathogen insect. It has led to their ban in France in 2019. However, all alternatives were less effective and more harmful. No alternative practices such as organic farming provide a convicting alternative. Such limitations, called "technical dead end" since no proven solution providing similar productivity with less environmental impact is yet available.

Complexity will be fundamental to increase resilience, mitigate negative impacts while producing enough food. Such complexity will require many experiments to build sufficient knowledge. In the case of sugar beet, scientists envision that the alternative will be a combination of genetic and farm management to repel the aphid and attract him to other parts of the field, to have a different crop that limits the spread of the pathogen, have varieties which are resistant to the viruses. Strategies need to adapt since insects and diseases will probably find ways to avoid too simple solutions. For instance, the rust resistance that made the wheat cultivars successful for the Green Revolution was based on two genes. At the end of the '90s, new strains of rust have bypassed the defence mechanism,

making the varieties less relevant. The pace of experiments for agriculture must increase to find more complex and dynamic strategies. Experiments have always been key in agriculture, but the dynamic nature of crops and the associated environment and the inherent difficulty in evaluating outputs make progress slow.

1.1.5 Towards genomic selection

Since the 2000s, a new era of crop breeding has emerged thanks to a better description and understanding of DNA. Instead of crossing two parents in the hope of adding few genes of interest into the genes pool of a high yield cultivar, the genome is precisely characterized with methods such as SNP (single-nucleotide polymorphism) or SSR (simple sequence repeat). A SNP is "a single base pair mutation at a specific locus" [34]. Genotyping is today cheap and accessible. Such innovations are now accessible thanks to the low cost of sequencing techniques that describe the genome : rice [35], barley [36], maize [37] , soybean ([38] and more recently wheat [39]. Association between genome and phenotype can be measured with "Genomic Wide Association Study" (GWAS). The method finds "Quantitative Trait Locus", a specific region of a chromosome responsible for the variation of a phenotypic trait by observing the concurrent changes in SNP distribution and the phenome. Models that predict phenotypic traits forgiven a genome are developed. The more scientists can understand the genome's role, and the faster new varieties can be created. The study of transcriptomics is also of interest by understanding the expression of a gene through RNA sequencing. It has already led to applications such as development stage control by silencing or overexpressing some genes [40], [41]. As illustrated in Figure 4, the introduction of genomic variation can be done not only by selection (Conventional plant breeding) but also with induced mutation thanks to chemicals of radiation (plant mutation breeding). Foreign genes can directly be introduced into a host (Plant transgenic breeding), with bacteria such as Agrobacterium tumefaciens commonly used inbreeding. The introduction of faster, more precise techniques known as genome editing enable precise modifications of few bases. The most known strategy is the CRISPR/Cas9 [42] which allow determining specific targeting sequence for the new DNA addition. It is a challenging exercise as genes can have interactive effects: genes that induce a disease's resistance can trigger a physiological process that can lead to a loss of yield.

1.2 High-Throughput phenotyping is a key tool for this transition

1.2.1 The bottleneck of the digital and genomic revolution: the phenotyping problem for agricultural experiments.

Experimenting in agriculture is central to improving knowledge, evaluating management practices' efficiency by comparing modalities, evaluating new accessions and cultivars, and calibrating crop models. The field experiment is therefore usually positioned such as the soil is homogeneous, the area around the trial is protected with a large buffer zone, and all modalities or genotypes are tested on a small piece of land with a fixed number of rows and a fixed length. Each modality is often replicated on several elementary experimental units, which are called a "microplot". Conventional plant breeding takes 8 to 10 years. This process is done in several steps: first individual plants are grown in-door to get as many crosses as possible. Then, given a set of criteria, the best performing genotypes are kept and used for a new test with more space. Finally, candidate genotypes are tested in a set of different environments. This process is usually done by commercial companies, universities or non-profit organizations such as CIMMYT. In France, the quality of a genotype is also independently tested on a national network to provide independent feedback to farmers. At this end of this process, the selected genotype can be registered as a variety, or cultivar, if it demonstrates a significant improvement compared to a set of reference varieties. At each

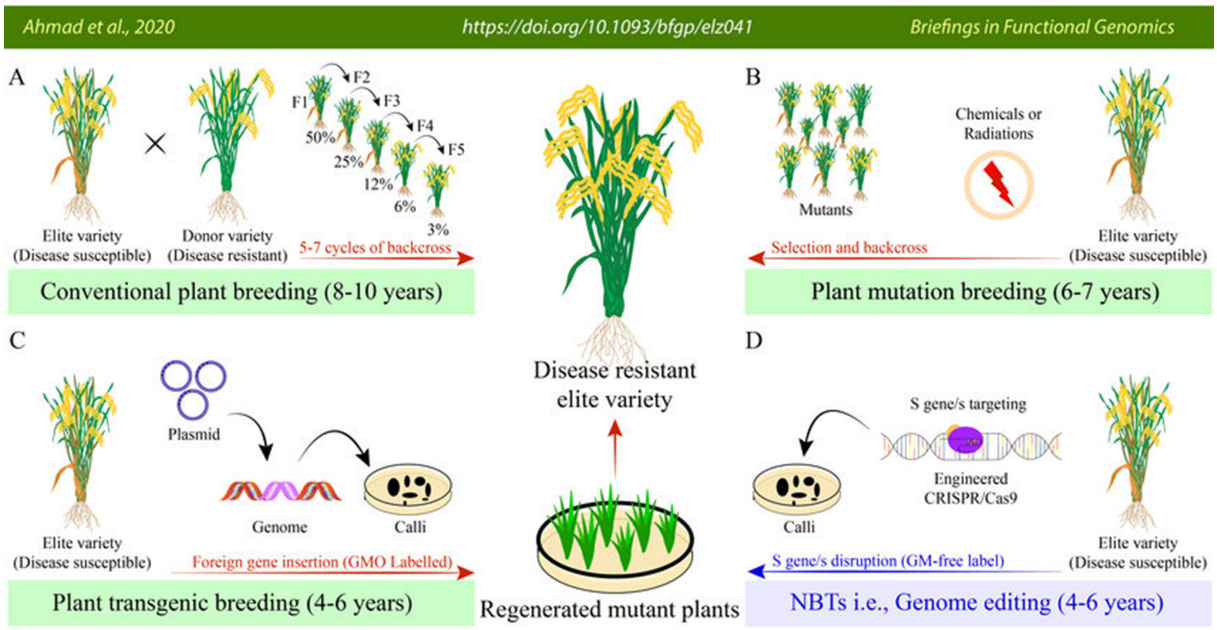


Figure 4. The different plant breeding strategies. Reproduction from [43] with author's authorization.

step of these experiments, many measurements are involved in assessing the performances. The exact process is applied to agronomical experiments to evaluate the efficiency of inputs, new cropping systems or other agricultural innovations.

Measuring the performances is still mostly done manually by field workers. Measurements can be either destructive or non-destructive. Destructive measurements are made by gathering the plants in a given sampling area and are used to quantify several traits, including the dimensions of the organs, their dry biomass, or the nitrogen content. Destructive measurements can be repeated only a few times during the season, considering the size of the microplot and sampling area required each time. Conversely, non-destructive measurements can be replicated multiple times during the season. It is used to measure the evolution of architectural traits such as canopy height or plant and organ counting. Visual notations are also used to evaluate some traits but are prone to variability among the different field workers. When relying on the field workers, another challenge arises around critical development stages: the number of activities is so high that not all needed measurements can be done, leading to a long day of work for field workers and incomplete data for agronomists. These methods were exploited during the Green Revolution for plant phenotyping. However, this currently constitutes a substantial limitation for the current genetic revolution based on high-throughput genotyping. Therefore, Plant phenotyping appears to be now the bottleneck of genetic improvement.

High throughput plant phenotyping (HTPP) is a recent field of research exploiting the most recent technological advances in vectors, sensors and interpretation algorithms to estimate traits characterizing crop architecture, physiology or disease symptoms. Additionally to its interest in plant breeding, the same technologies can be adapted to agronomical experiments and for farmers' decision-making. Before reviewing the current trends in HTPP, we first define few critical terms following the glossary proposed by [44].

- ***Genotype***: The genotype of an organism is a set of inherited instructions carried within its genetic code.
- ***Trait***: A trait is a quantitative or qualitative characteristic of an individual resulting from the expression of its genome in a given environment. For plant phenotyping, it can be determined at the plant scale or at population scale. The term "phene" is also used.

- ***Variable***: A variable is an estimate of the trait resulting from the application of an interpretation method on sensor measurements.
- ***Phenotype***: the phenotype is the set of observable traits of an organism. It covers morphology, development, cellular, biochemical, or physiological properties.
- ***Phenome***: the phenome is the set of all possible phenotypes for a plant or a population.
- Phenomics: The study of the phenome and its relationship with the genome, the transcripts, the proteins or metabolites.
- ***Ideotype***: A crop ideotype is an ideal collection of traits which optimize crop performance to a specific end-use in a particular environment and crop management.

The role of HTTP in plant science is to develop automatic systems that measure traits. The phenotyping systems are made of three main components: a vector that can move the plants to the sensor or the sensor to the plants, a sensor that records a signal from the plant, and an interpretation algorithm that transforms the signal into a trait estimate. These systems are operated over experimental installations with a range of levels of control of the environmental conditions.

A *phenotyping* process can be defined as "high-throughput" if the acquisition for one microplot is a magnitude faster than a manual measurement [45]. However, the measurements' cost efficiency and accuracy need also to be accounted for when compared with traditional phenotyping methods. In the next section, different HTPP systems will be presented, the produced traits will be analyzed, and finally, the current interpretation algorithms will be discussed. While HTPP can be viewed as one engineering science, its multidisciplinary nature (robotics, mechatronics, statistics, computer vision, machine learning, artificial intelligence, modelling, agronomy, ecophysiology, botany) allows the emergence of "Plant Phenomics": a science that relates the phenome with the genome, the transcripts, the proteins, and the metabolites.

1.2.2 The several traits accessible from HTPP

The measurement of traits with HTPP produce variables: a discrete or continuous quantity that quantifies the traits that can then be used to compare different individuals. Traits availability and the precision of the associated variables depend both on sensors' technological development and interpretation methods maturity. Four categories of traits are proposed and summarized in Figure 5:

- ***State traits*** which are measured at a specific time and are intrinsic properties of the crop/plant/organ
- ***Dynamic traits*** which are based on repeated estimates of state traits over a specific period but do not need additional inputs except a time scale
- ***Radiative traits*** which are measuring the flux reflected or emitted by the plant that will depend both on plant state and on the illumination and view conditions
- ***Functional traits*** that describe some processes and that need additional information about the environmental conditions.

States traits

The states traits can be measured directly on the canopy and take exactly one value at a precise time. They can be categorized into three groups:

- ***Biophysical traits*** describe the structural and morphological characteristics of canopies, plants, or organs. The aerial parts of the plant that are more easily accessible are often described: at the canopy level focus is generally on Green Fraction, Green Area index, Leaf orientation, plant height, lodging, plant density, ear density.
- ***Biochemical traits*** provide information on the plant biochemistry. For instance, the chlorophyll content of a leaf or a canopy is an important biochemical trait that determines the photosynthetic potential.

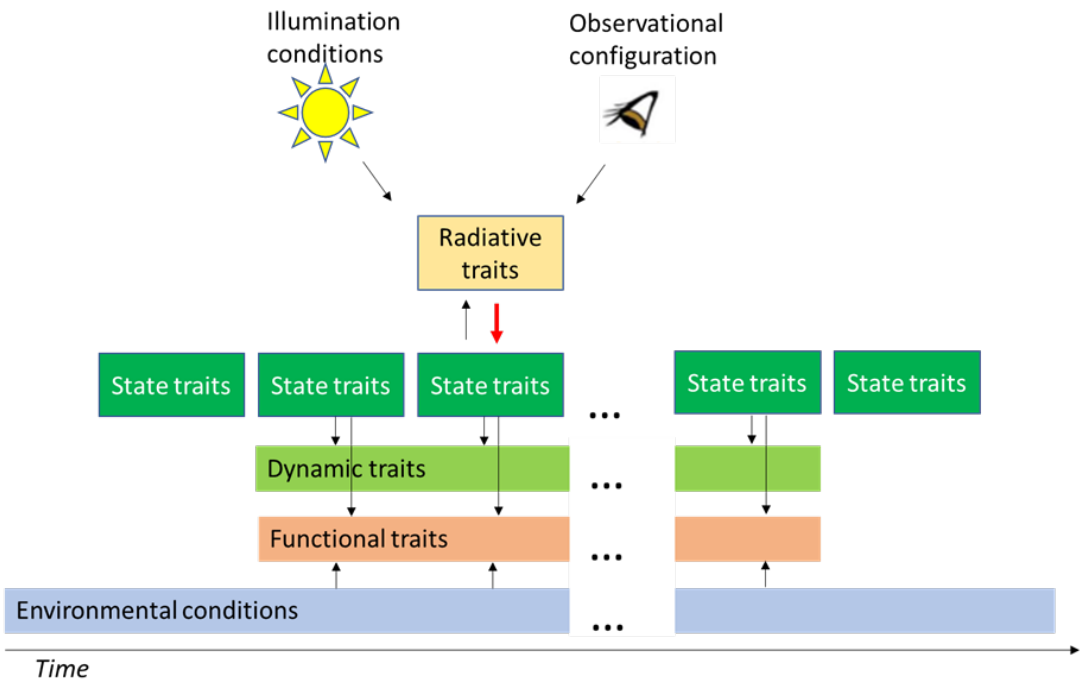


Figure 5. The different traits in phenotyping

- ***Sanitary traits*** provide information on the plant health. For example, the detection of disease symptoms, the determination of contaminated parts of the plant are sanitary traits that are often derived from visual scoring of the symptoms severity made by human experts.

Dynamic traits

Dynamic traits are based on repeated observations of state traits. Popular dynamic traits among breeders are the early vigour or the stay-green. The early vigour is the plant or canopy growth speed, while the stay green is the plant or senescent canopy rate. Specific characteristics of the plant architecture plasticity are evaluated dynamically, such as the leaf rolling. Phenological traits are also dynamic traits that are measured by detecting qualitative changes in plant morphology. In wheat, tillering, stem elongation and heading or flowering are evaluated by monitoring biophysical traits such as plant height or wheat head density.

Radiative traits Radiative traits result from the interaction between light and canopies, plants or organs. They thus depend both on the view and illumination conditions and canopy, plants, or organ state. Canopy reflectance or organ radiance are low-level radiative traits commonly measured and then transformed into higher-level traits such as vegetation indices. The fraction of intercepted photosynthetically active radiation is one of the main radiative traits computed from canopy structure and the illumination conditions. Brightness temperature is also a low-level radiative trait that depends both on canopy state and illumination (and environmental) conditions. Chlorophyll fluorescence is related to photosynthesis activity and depends on illumination (and environmental) conditions and canopy state.

Functional traits Several definitions of functional traits exist in the literature due to the concept being explored in the context of plant phenotyping and ecology. Caruso [46] proposes that "functional traits are generally considered aspects of plant phenotypes that influence growth, survival, and reproduction by mediating interactions with the biotic and abiotic environment". We propose to define functional traits as traits describing canopy, plants, or organ reactions to the environment. Since they account explicitly for the environmental

conditions on some processes, they are expected to be less sensitive to some environmental factors. They will therefore be more heritable than most of the other traits. Efficiency traits are commonly used functional traits that evaluates the efficiency with which elements are used by the plant to grow. They include the radiation (RUE), water (WUE) and nitrogen (NUE) use efficiencies.

A more exploratory approach for functional phenotyping is to adjust the parameters of crop growth models to match the dynamics of the state variables. Blancon [47] proposed to calibrate a simple maize growth model [48], [49] using Leaf Area Index calculated from UAV RGB imagery to retrieve parameters that were more heritable than the LAI itself. Recently, Shouyang Liu [50] proposed to use a 3D model to retrieve few parameters of the ADEL-wheat model. The advantage of HTPP phenotyping is also to propose traits that can be use more frequently by crop models such as CHN [51], Apsim [52][53][54] or STICS [55][56][57], which represent our current knowledge on plant physiology. Advanced functional traits are meant to breed on specific processes that allow a better understanding of the reaction of the crop under a wide range of environmental conditions. As a result, it is possible to find an optimal set of functional traits that will define an ideotype for given environmental conditions [58].

1.2.3 Phenotyping installations to control/describe the environment and measure traits

The several types of installations

Experimental installations, also called field stations or platforms, are the facilities where the experiments take place. One or more trials can be conducted in such installations. A platform is equipped to describe the environmental growing conditions and the several traits of interest. The environmental conditions can be fully controlled as in ecotrons [59], or only some environmental factors are controlled/manipulated as in field conditions. HTPP installations have a high degree of automatization both for measuring the environmental conditions and the plant traits. Installations can be very diverse, and some examples are shown in figure 6:

- ***Low control / Low detail*** The most basic installations, often used for the study of agricultural practices (e.g. inputs, farm system etc...). It can be as simple as a farmer's field with few large microplots.
- ***Low control / High detail*** A high-cost installations which does not control the environment is used to study the performances in real conditions with high temporal and/or spatial resolution and equipped with numerous instruments. The Arvalis station in Gréoux Les Bains, or INRAe in Auzeville, both in France belong to this category since they are equipped with a Phenomobile, an automatic rover system. Such installations are used in the final stage of the crop selection and cultivar evaluation.
- ***High control / Low detail*** A low-cost installation that controls the environment is used to measure few key traits of plants to determine the impact to specific environmental conditions. It is still used to measure the plant transpiration for instance.
- ***High control / High detail*** An installation with high control and high throughput systems to measure accurately many traits under a range of fully controlled environmental conditions.

The phenotyping systems used to measure traits

Under field conditions, a vehicle, called the vector, moves the sensors that record the signals emitted or reflected by the plants. The combination of a vector and sensors determine a specific temporal and spatial resolution and coverage, as presented in figure 7. Satellites such as a Sentinel or Planet provide field data a few times per week everywhere globally, at the expense of a metric to decametric spatial resolution. Gantries and Robots are the most advanced vectors used for phenotyping purposes. These systems can be automatic (Phenomobile [60], FIP[61], Rothamsted Field Scanalyzer[62]) or modified agricultural machinery

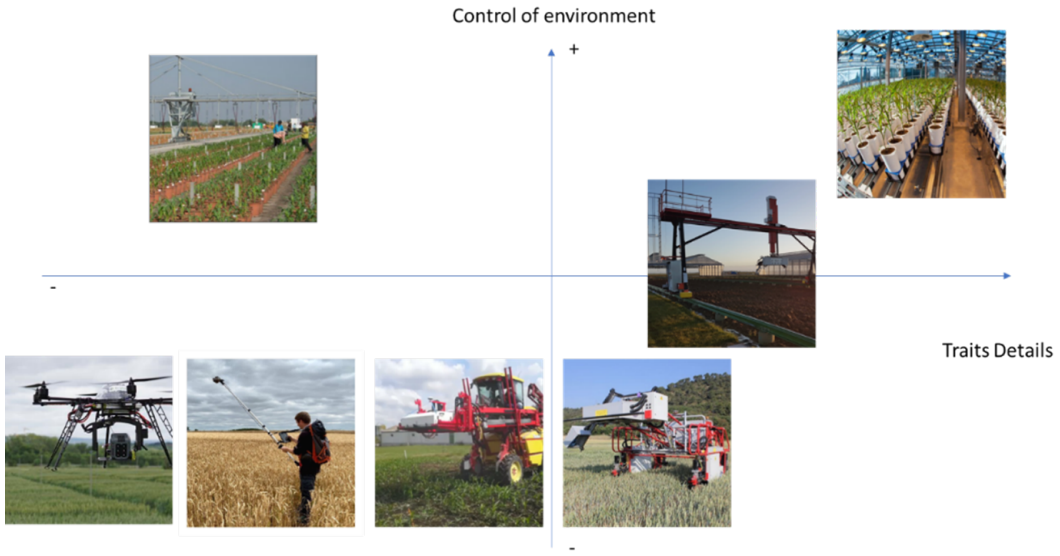


Figure 6. : Classification of the experimental installations defined by the targeted levels of trait details and control/description of the environmental conditions.

driven by a human such as a tractor [63]. An important feature is the presence of an active illumination or not. These vectors can "scan" the agricultural experiment every 3-7 days with a very high spatial resolution and a flexible viewing geometry. A cheaper alternative appeared: the UAV, which is an efficient solution for plant phenotyping, achieving very high throughput with a high spatial resolution and numerous flights covering the whole crop cycle and acquired traits such as the green fraction [64], plant counting [65], [66] or height [67]. Hand-held devices or smartphones can also be used for phenotyping in the field, allowing some key traits to be measured. Such devices are attractive to conduct experimentation in farmers' fields and are already used for decision making for farmers: apps such as Xarvio scouting or PlantVillage Nuru[68] provides automatic advice based on a smartphone photo. Finally, field IoT (Internet of Things), which are connected cameras fixed on a boom, can monitor daily the growing stages [69] and disease development [70]. However, the coverage of a single system is minor, in the order of few square meters.

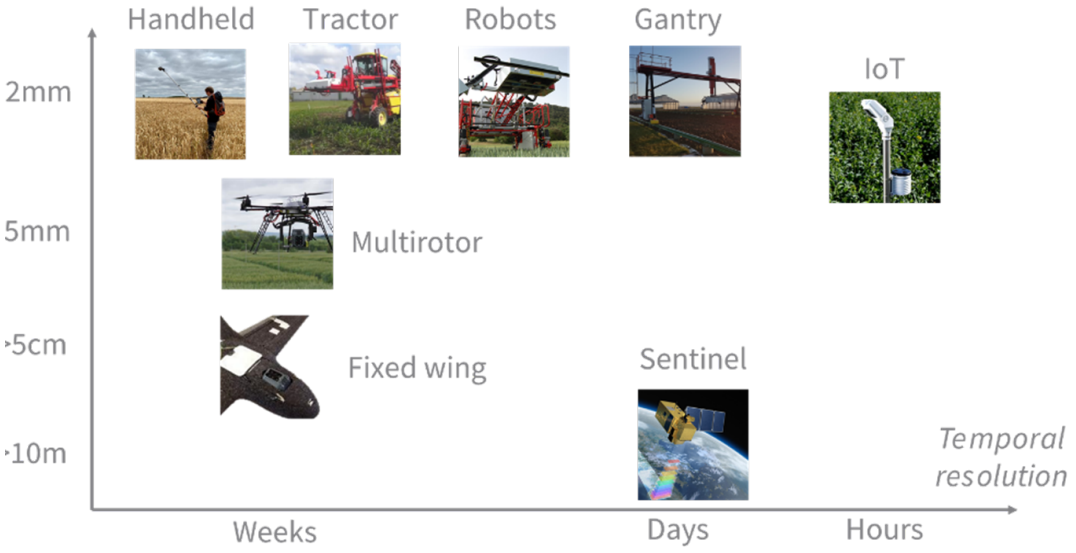


Figure 7. : Presentation of the different vectors.

The choice of the sensors is constrained by the vector and is generally a trade-off between cost, possible payload and energy consumption, spectral resolution, and spatial resolution. RGB imagery is the cheapest way to acquire data, with an unmatched resolution of millions of pixels and is widely used aboard all vectors. Recent drones can be equipped with a 100Mpx camera. Multispectral and hyperspectral imaging systems are popular for phenotyping applications, as these sensors can retrieve information on the biochemical composition of plants. However, the gain in spectral richness is achieved at the expense of spatial resolution. Vectors are also often equipped with a lidar system to reconstruct the 3D structure of the canopy finely.

While the HTPP systems are designed to increase the throughput with the possibility of monitoring crop growth with a high-revisit frequency, they also improve the accuracy with which the traits are estimated and allow access to new traits not measurable with the traditional techniques. However, the large volume of data generated and the complexity of interpreting the recorded information currently constitute the bottleneck of phenotyping. In the following, we will focus on this third component of HTPP systems: data interpretation.

1.2.4 From remote sensing interpretation methods to deep learning ones

Trait	Description	Category	Unit	Method	Phenomobile (RGB)	Handheld (RGB)	UAV (RGB)	UAV (Multispectral)
GreenFr	Green Cover Fraction	State - Biophysical	Unitless [0:1]	Segmentation with DL or SVM	✓	✓	✓	✓
GAI	Green Area Index	State - Biophysical	Unitless (% of observed area)	radiative transfer model inversion	✓	✓		✓
ALA	Average Leaf Angle	State - Biophysical	deg	radiative transfer model inversion	✓	✓		✓
FIPAR	Fraction of intercepted PAR	Radiative	Unitless [0:1]	radiative transfer model inversion	✓	✓		
CropFr	Crop Cover Fraction	State - Biophysical	Unitless [0:1]	Segmentation with DL	✓	✓		
SenescentFr	Senescent Fraction of the crop	State - Biophysical	Unitless [0:1]	Segmentation with DL	✓	✓		
HeightMax	Maximum plant height	State - Biophysical	m	Photogrammetry	✓	✓	✓	✓
SpikesDensity	Density of spikes	State - Biophysical	number/m2	Detection with DL	✓	✓	∽	
PlantsDensity	Density of plants	State - Biophysical	number/m2	Detection with DL			∽	
NDVI	NDVI	State - Biophysical	Unitless	Vegetation Index				✓
Clgr	Cl green @ nadir	State - Biochemical	Unitless	Vegetation Index				✓
MTCI	MTCI @nadir	State - Biochemical	Unitless	Vegetation Index				✓
LCC	Leaf Chlorophyll Content	State - Biochemical	microgram of Chl/cm2 of leaf	radiative transfer model inversion	∽			✓
CCC	Canopy Chlorophyll Content	State - Biochemical	g of Chl/m2 at ground level	radiative transfer model inversion	∽			✓
LodgingScore	Plant Lodging	State - Biophysical	Unitless [0:1]	Photogrammetry			✓	
DiseaseFraction	Fraction of the crop (or organ) affected by a disease symptoms	State -Sanitary	Unitless [0:1]	Segmentation with DL	∽	∽		
Phenology	Growth Stage	Dynamic	Zadock scale	Classification with DL	∽	∽		

Table 1. Presentation of the available traits at Arvalis for production (✓) or in development (∽)

High throughput plant phenotyping is today an independent field with its journals and

methods, although it was created by an association of remote sensing scientists and geneticists. It strongly influenced the first algorithms used to transform raw data acquired by HTPP into valuable traits. The focus was first put on exploiting the spectral information with the use of vegetation indices that are simple combinations of reflectance or radiance in few spectral bands. For instance, NDVI that combines the red and near-infrared (NIR) bands is a good proxy of green vegetation. Several vegetations indices were proposed in the literature for hyperspectral, multispectral and even RGB sensors [71], [72]. Multispectral and hyperspectral cameras are popular to estimate the content of pigments used in photosynthesis, such as chlorophyll. However, most of these simple interpretation methods are based on empirical relationships with few traits. These relationships are generally established over a limited dataset, and great caution should be paid when applying them under conditions not represented in the training dataset.

Further, the simple vegetation indices may also lack causal relationships with the targeted trait, resulting in poor accuracy and robustness. Alternatively, physically-based approaches have been developed to get more robust estimates of a few critical structural and biochemical traits. The PROSAIL model [73] is a popular radiative transfer model that combines PROSPECT [74] that simulates leaf optical properties, and SAIL [75] that simulates bidirectional canopy reflectance. PROSPECT was used to estimate GF from the multispectral reflectance. When observing at 45° inclination, we can estimate the GAI (Green Area Index), the AIA (Average Inclination Angle) when combined with nadir observations, which can finally be used to compute FiPAR (Fraction Intercepted Photosynthetically Active Radiation), which summarizes the biophysical capacity of a plant to capture solar radiation. The FIPAR can then be used to calculate the light use efficiency if independent biomass measurements are available. However, PROSAIL is based on simple assumptions on canopy architecture that are well verified for crops such as wheat while being violated for other crops such as maize or sunflower.

Other instruments complement the RGB, multispectral or hyperspectral cameras: Infrared thermal cameras measure the surface temperature that indicates water stress under specific conditions. Active and passive fluorescence cameras have also been used to access the functioning of photosynthesis and other characteristics of the leaves. More recently, approaches based on exploiting the high-resolution imagery have been developed, especially using RGB cameras. Computer vision techniques [76] based on geometrical rules have successfully estimated several structural and morphological traits. Further structure-from-motion (SfM) methods applied on UAV images allows generating the surface elevation model from which the plant height can be derived [67]. Alternatively, stereoscopic and multi-vision methods and LiDAR also provides a 3D description of the non-occluded parts of the plants [77]. Such methods are widely used for in-door phenotyping to describe the architecture of the aerial parts [78], and roots of the plants [79]. Vegetation indices, physically-based methods, computer vision ones, or Deep Learning approaches are already able to produce an extensive range of traits for HTPP. Table 1 presents an example of the traits that are routinely produced at Arvalis.

High-resolution RGB imagery allows the exploitation of the texture and the shape of the objects in the image to extract a range of traits. Compared to the multispectral and hyperspectral imagery where the lower spatial resolution forces to interpret the signal concurrently for identifying the organs of interest and characterizing them to derive the targeted trait. It results in possible confounding effects, leading to generally to lower performances of the estimation. The interpretation of high-resolution RGB imagery has benefited for approximately five years from the advances in deep learning techniques. Several traits are now estimated using DL, including plant segmentation ([80], [81] and plant and organ detection [82], [83], disease classification and quantification ([84][85][86], biomass and yield prediction ([87], [88]). Deep learning is now considered as the state-of-the-art approach that handles a

large diversity of traits difficult or impossible to estimate with pixel-based classification and computer vision techniques [89].

1.3 Robust estimation of traits with deep learning

The development of DL models requires much human effort to prepare large datasets of labelled images used both for training the model and its evaluation. It is due to the empirical nature of the DL approach, inferring the predictions rules from a training dataset. It poses a problem of reliability of the predictions when applied to an extensive range of HTPP data. This section first presents the principles of Machine Learning and Deep Learning, emphasising the robustness problem.

1.3.1 Basic of Deep Learning in computer vision

A brief history of recent progress in computer vision

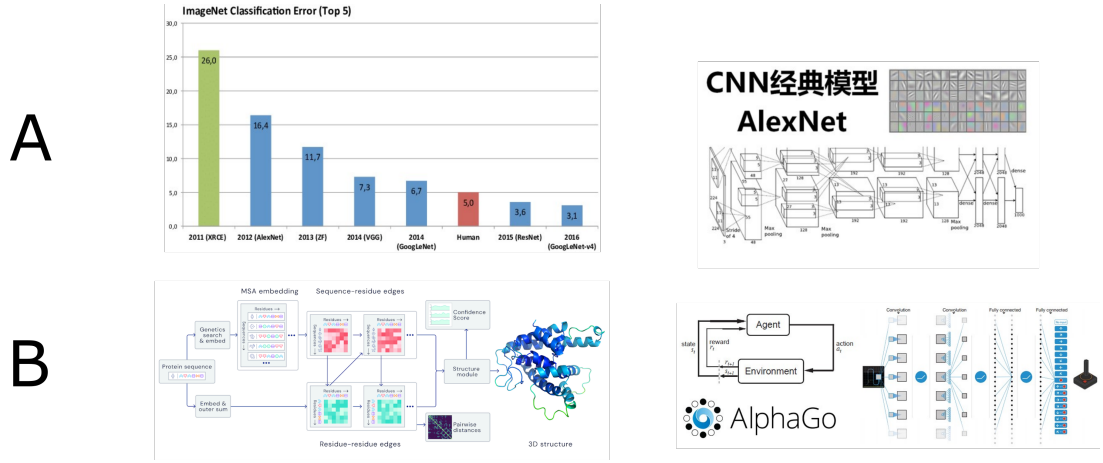


Figure 8. Illustration of some of the most important breakthroughs made by Deep Learning

Deep Learning is a subfield of Machine Learning (ML) and is based on multiple processing layers to extract features and solve a large range of tasks. It achieves state-of-the-art performances in a wide range of applications including speech recognition [90] , visual object recognition [91][92][93][94], localization [95][96][97] , segmentation [98], language translation [99], and plant phenotyping as previously described. A neural network is still the state of the art of ML for relatively simple problems. It is based on the use of a network of neurons. A neuron is an elementary processing unit that receives several inputs, which are transformed into one output. The most simple form of a neural network is the perceptron [100] which is made of one neuron. However, neural networks are generally organized in several layers of neurons and are called Multi-layer perceptron (MLP) [101]. The neuron's input can be the raw input data or the outputs of the previous layer.

Although neural networks were already described in the '60s, DL was discovered supposedly in 1989 [102] or 1995 [103], during LeCun stay at the ATI&T labs. The first DL model was able to recognize cursive letters on mails thanks to few layers of convolutions learned with backpropagation [104]. DL was, however, not very popular at this period as compared to other ML algorithms. The rise of DL started with the winning of AlexNet [94] on the ImageNet [105] large scale visual competition: it beats XRCE ([106]) by almost 10 points (from an error of 26% to 16% as shown on figure 8). From this date, no alternative approaches to DL won any major challenges in computer vision, almost ten years later.

These successes have spread to other domains such as language processing which also use DL intensively since the introduction of the attention mechanism by Bengio et al. [107]. Figure 8B presents two unbelievable breakthroughs made by the DeepMind company: Alphazero [108] which beats a human on the chess and go games without looking at historical games, and AlphaFold [109] which converts the 3D structure of proteins from their 2D sequence. This problem previously required a lot of human inputs with the program FoldIt [110]. From a theoretical point of view, the main breakthroughs were theorized well before the emergence of AlexNet: convolution, backpropagation, neural networks were already known but not widely used. Practical factors are contributed to the development of DL for many applications:

- The release of CUDA [111] in 2007 by NVIDIA allowed researchers to use GPU with languages such as C, C++ or Fortran. DL requires a lot of computer power, but most of the operations can be parallelized. A GPU can be viewed as a set of thousands of low-frequency CPUs. AlexNet was one of the first networks implemented on GPU, allowing for more layers and training on more data.
- Large labelled datasets were available starting 2009, with ImageNet release [105]. It was on the first time that 100 million images were used to train a network. Later, other datasets such as the MS COCO [112] were crucial to train detectors. Such benchmarks helped researchers to focus on improving the architecture and increase the performances of the DL models.
- The quality of DL libraries also has helped to democratize. Caffe, written in C++, was one of the first libraries that could be leveraged to adapt DL networks for new domains, alongside Torch, written in Lua. The emergence of Tensorflow, which had a python API, also helped a lot. Today, multiple frameworks exist. Pytorch is popular among researchers while more beginner-friendly yet powerful libraries exist, such as Keras. Figure 9 presents few popular libraries, sorted by their level of ease.

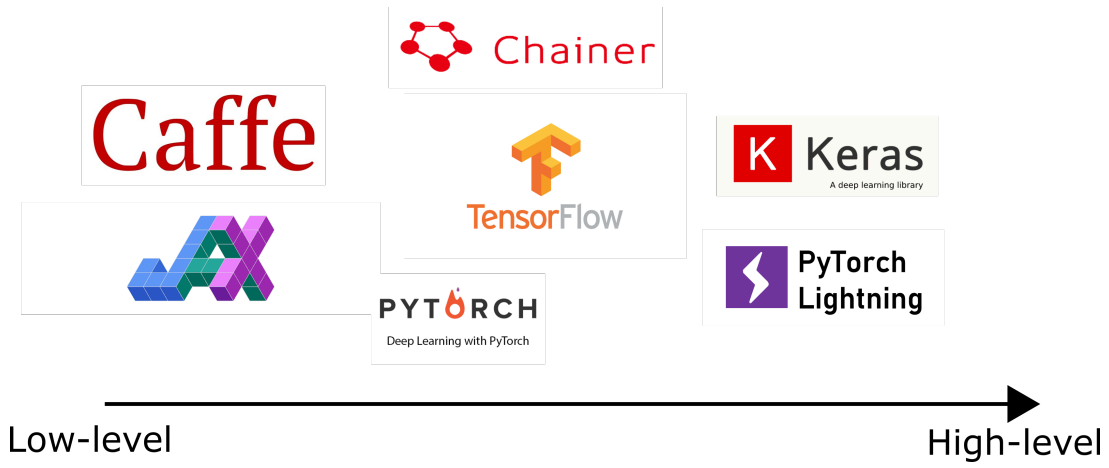


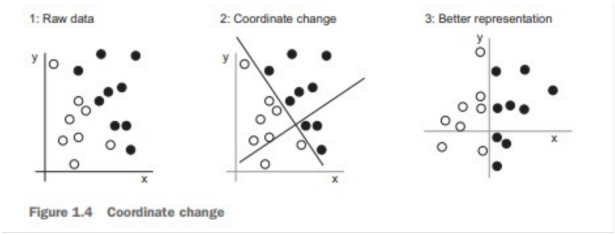
Figure 9. History of the Deep Learning packages used to interact with the GPU

The combination of industrial innovations was possible thanks to large companies such as Google, Facebook or Microsoft. This support materializes in different aspects: production of essential open-source codes, including Plant Phenotyping and open-source publication.

Deep Learning is about learning a better representation

In the traditional ML approach, an algorithm extracts features from an object of interest (image, text, audio file), and a classifier is trained to decide on these features. Choosing the right set of features is then critical to get high performances. In DL, the representation, i.e. features extraction, and the classification processes, are trained jointly, so no human interactions are needed to find the best features for a given problem. It is illustrated in figure

20 for image segmentation based on pixel colours. Instead of choosing a set of vegetation indices or changing the colour space to train a Support Vector Machine (SVM), DL can learn a directly used representation to segment an image. The change of representation helps to simplify the problem as we show it in example A (Figure 10): in this example, the better representation allowed by the simple rotation of the space reduces the complexity of the classification that can be done only on the x-axis. It can be viewed as a big and non-linear principal components analysis. DL methods are the most powerful way to learn a representation, but such an approach is not new in ML. Mairal and Ponce introduced the concept of "supervised dictionary learning " [113] where a set of wavelets are learned to compress the data.



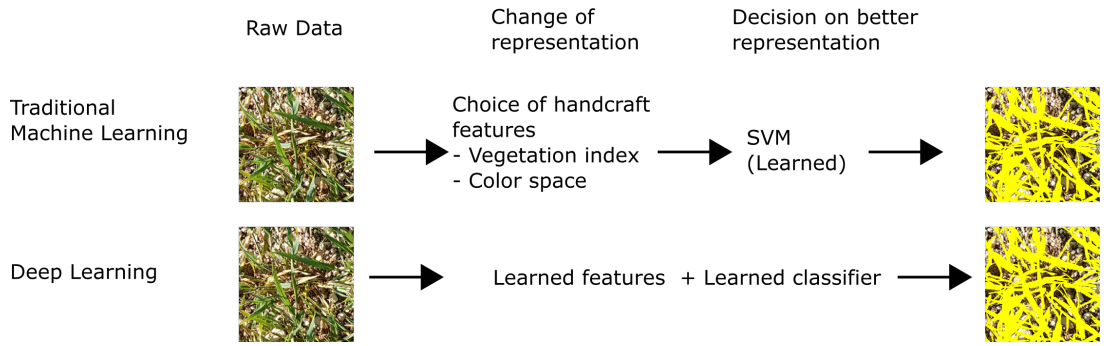


Figure 10. Comparison of change of representation in machine learning and in Deep Learning (Credit: François Chollet, Deep Learning with Python)

Fundamental blocks of convolutional networks

As described in 1.3.1, the popularity of DL is due to the emergence of convolutional neural networks (CNN), which are only one subclass of possible DL algorithms. One CNN is made of few key blocks that help to learn a better representation. In Figure 11, we can differentiate the two components of a DL model: the encoder, or backbone, that extracts the features, and the task solver, which is usually a stack of fully connected neurons. The encoder is made of successive convolutional layers. Convolution is the multiplication of one part of the matrix (i.e. one part of the image) by a weights matrix. All elements are then summed and affected to a pixel of the output. Then the convolution works as a sliding window to get the value for the other pixels. We present an example 11 of one convolution output (ResNet-50, first block) on a wheat image. It shows that the role of the convolution is to activate or de-activate some specific parts of the image. In our example, the leaves are activated in some dimensions but not in others. This process is repeated several times: the number of dimensions increases at each iteration while the spatial resolution decreases. The deeper we go, the more complex the representation of the image is.

To reduce the spatial resolution without learning additional parameters, one can also use a pooling layer, which sums or averages the values of a group of adjacent pixels to a unique value. This operation does not change the dimension but reduces the spatial resolution. The output is a unique features vector, usually between 1024 and 2048 dimensions, which are then used by a classifier. The features vector is meant to discriminate specific parts of

the image. A last important building block is the activation function, which is usually used on the convolution output. The activation layer, which needs to be differentiable, helps the gradient not to vanish and brings useful non-linearity to the network. Usual functions are presented in figure 12, but it is a research area by itself! All CNN models such as AlexNet [94], VGG-16 [114] , ResNet-50 [115], and EfficientNet [116] propose variations of the described architecture. Some works are conducted with AutoML [117] to learn the optimal architecture instead of handcrafting it.

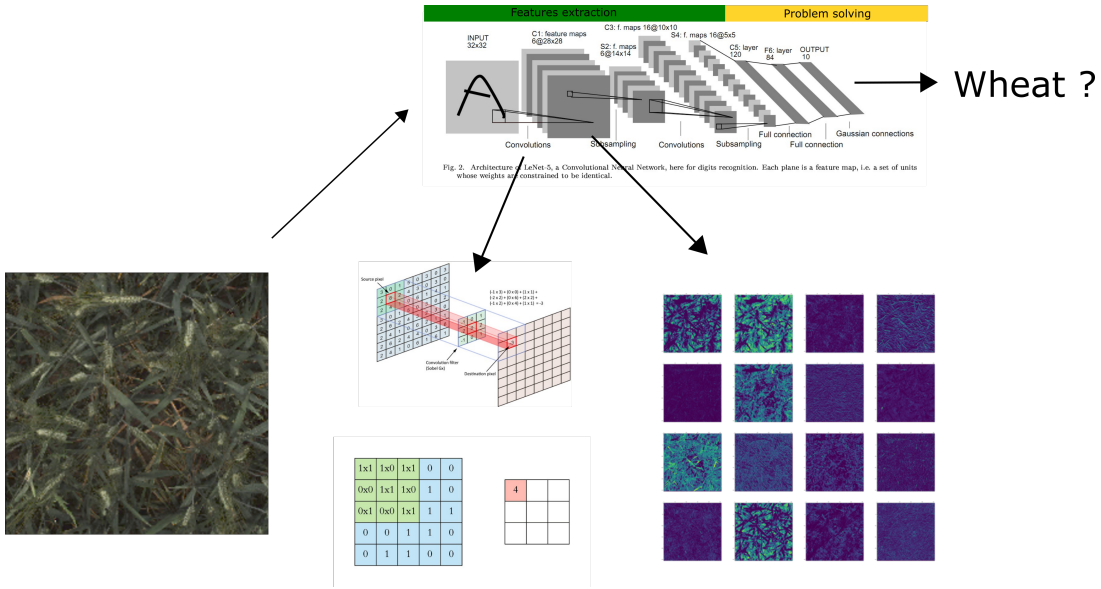


Figure 11. Illustration of the processing of a RGB image by a CNN

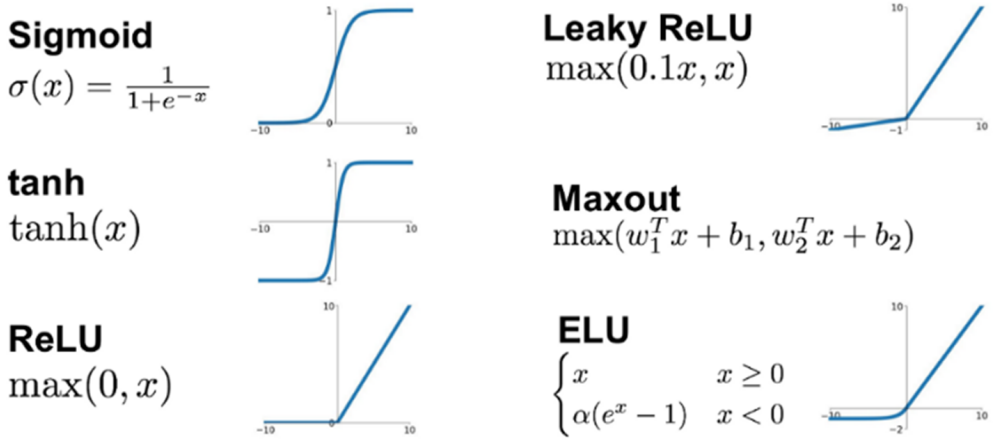


Figure 12. Presentation of different activations functions

Training a Deep Learning network right

The architecture of CNN models is not the only reason for their success, but the learning process, called *training*, is important to tune optimal weights for the convolution. It is done by using a "loss function". During the training process, presented in figure 13, each data point (for instance, an image) is presented to the network, which will provide an output (for instance, the class "wheat" or "maize"). This output is compared to the labelled target ("Wheat"), and the loss function computes a penalty. The goal of the training is to minimize this loss by changing weights values. The derivative of the loss function is used to update the weights and minimize the loss. The backpropagation gradient, or "backward propagation of

errors", introduced by Rumelhart and Hinton [104], helps to efficiently evaluate the gradient of error at each layer of the network. The *optimizer* is the algorithm that updates the weight given the gradient value. A simple optimizer such as the stochastic gradient descent (SGD) updates the weights by subtracting the gradient multiplied by a *learning rate* value. The learning rate can be interpreted as the sensitivity to new data: a high learning rate will make the network prone to overfit, while a low learning rate prevents the network from learning. The loss function can be adapted to plant phenotyping scenarios to encode expert knowledge. For instance, in the case of multi-species vegetation classification, one can set the model to predict the class "vegetation" AND "wheat" or "maize" and set a penalty if the algorithm attributes the class "background' and "wheat" to the same pixel. It is an area of research that is more probably accessible to the plant phenotyping community.

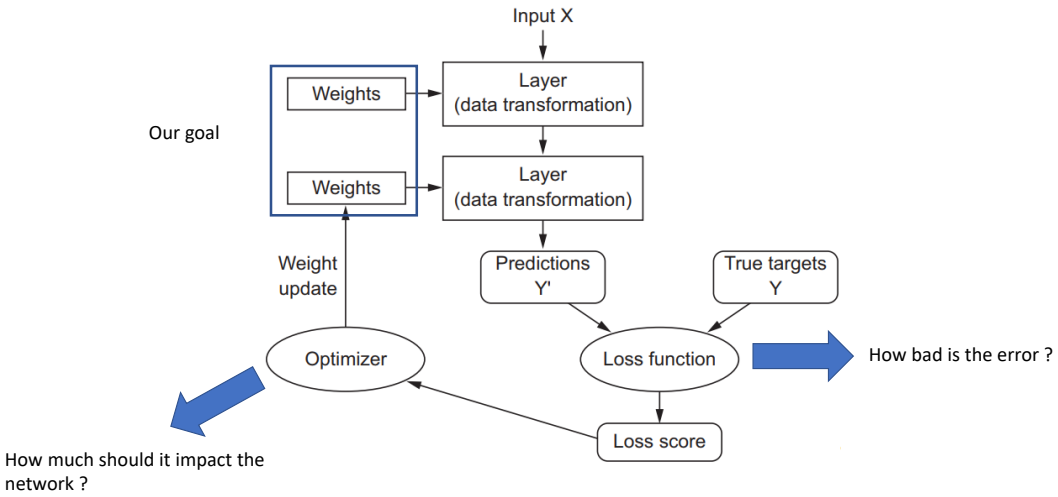


Figure 13. Illustration of the optimization of a Deep Learning Network (credit: François Chollet, Deep Learning with Python)

Training from scratch a DL model on a very large database is time consuming and costly. The completion of 90-epoch ImageNet-1k training with ResNet-50 on a NVIDIA M40 GPU takes 14 days. [118]) and costs 1.2 M USD to train it in 24 minutes. However, it can be optimized to 18 minutes with an associated cost of 40 USD [119]). It is then not always accessible for everybody to train a state-of-the-art detector. Also, not everybody has access to large datasets to train a network for their application. However, a popular technique called *transfer learning* helps to solve this issue. The learned representation of models trained on large datasets can be re-used for other problems, even with smaller datasets. The process is described in figure 14: the classifier can be removed while keeping the encoder (*trained convolutional base*), and one can append a new classifier on top of the representation. At the beginning of the training process, the error (*loss*) will usually be high. Therefore, we *freeze* the encoder so we do not lose the representation because of the first step, and we let the classifier train on top of good features.

Few model architecture have allowed the large varieties of algorithms in 2.4

DL models can be designed to solve more complex tasks than classification: they can use outputs larger than an n-dimension vector, and the loss function can be adapted. It is important to underline that most traits in plant phenotyping will be based on classic architecture and do not require designing a specific architecture. However, conferences such as the Computer Vision Problems in Plant Phenotyping (CVPPP, CVPPA) promotes plant

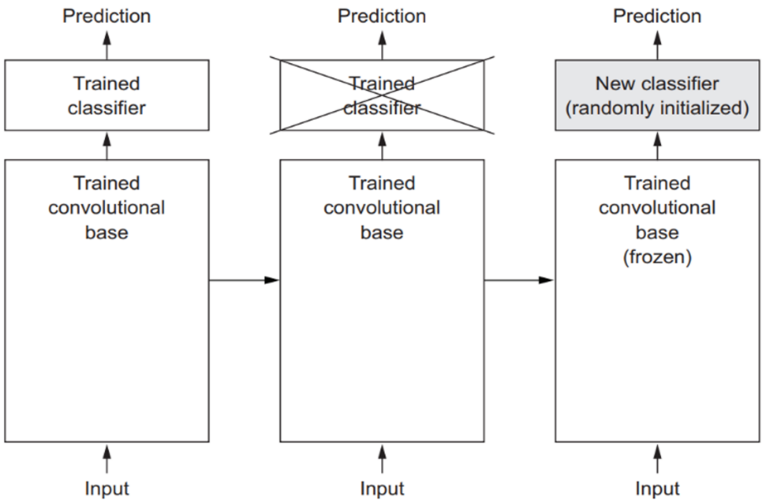


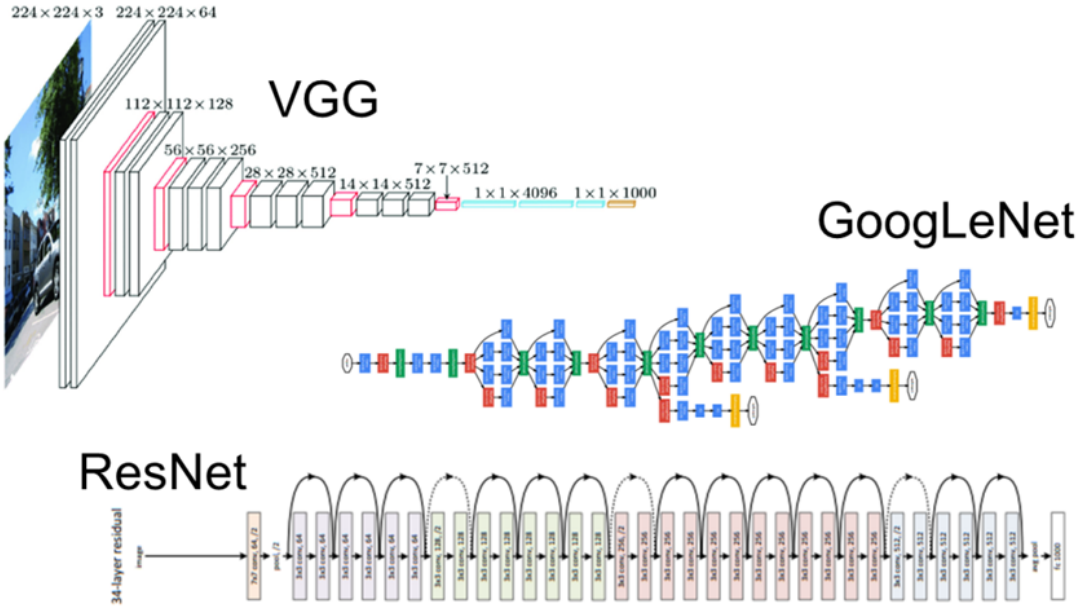
Figure 14. Presentation of a transfer learning process (Credit: François Chollet, Deep Learning with Python)

phenotyping as an area of research for Computer Scientists. Despite the vast number of models published every week, few stand above the crowd with time. Classification networks (figure 15a) are widely explored with architecture such as AlexNet, VGG, GoogLeNet which is very deep, ResNet [115], which introduces the concept of skip connection which concatenates the output of the convolution with its input, so information is better kept within a very deep network and does not vanish. Segmentation(figure 15b usually relies on the encoder-decoder architecture, with model such as U-Net [98]. After compression of the information to a very deep representation, up-convolution is used to restore the original resolution. Skip connection between the encoder and decoder is essential to share information of layers with the same resolution. The DeepLab [120] architecture relies on the "atrous convolution" to capture multi-scale information for segmentation usage. Such convolution allows getting a large field of view without increasing the number of parameters. Other encoder-decoder architectures such as Feature Pyramidal Network have been proposed for the segmentation process. A pre-trained classifier can use to initialize the weight of the encoder.

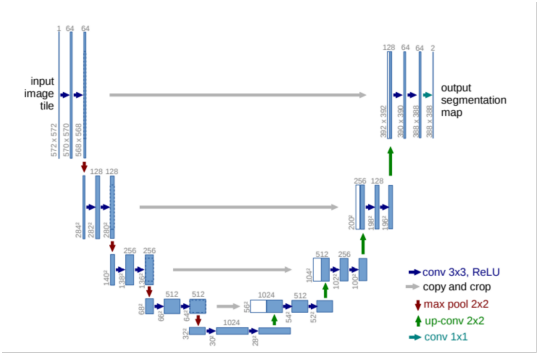
Encoders are also used for object detection, which can be separated into two families (figure 15c. One-Stage encoder which regroups SSD [121], YOLO [122], Mobilenet [123] and directly regresses the coordinates of the bounding boxes with their corresponding classes based on the encoded representation. A two-stage object detector will use the encoder to train a region proposal network (RPN) to propose possible objects. Then each object is extracted with its features with the Region of Interest pooling layer (ROI) and will be finally assigned to a class. One stage detector is usually faster to compute but less robust. Such object detector can be applied to instance segmentation, which retrieves not only a box around an object but also its mask (figure 16). Finding the best architecture can then be a challenge when designing a method to evaluate a trait. Models such as ResNet, Faster-RCNN or U-Net were designed between 2015 and 2017 and are still intensively used because training tricks or new layers can improve their robustness. These models are readily available in many DL libraries. The theory of DL is still moving, and cascades of new models are proposed every week. Convolutional neural networks are starting to be less central in the performance of DL.

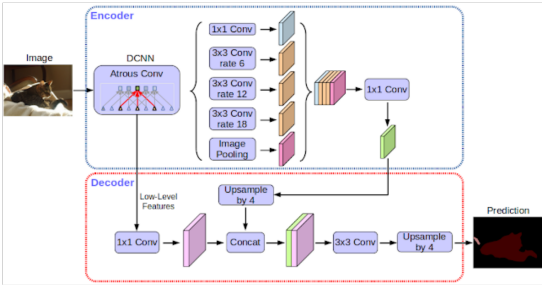
An explosion of alternatives to CNN

Convolutional Neural Network made it possible to learn DL features with neural networks by reducing the number of parameters compared to a multi-layer perceptron and is a significant block of the algorithms used in this PhD which focus on RGB imagery. However,



(a) Presentation of classification architectures

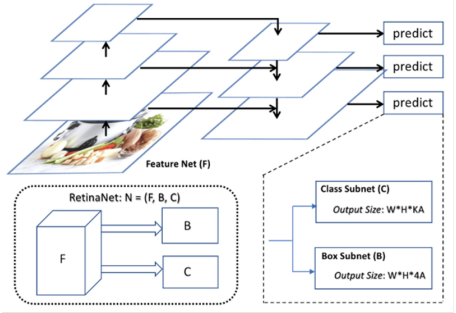


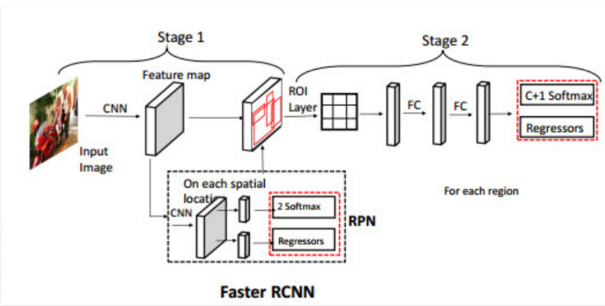


A- UNet Architeture (encoder-decoder)

B- DeepLab Architecture (Atrous convolution)

(b) Presentation of segmentation architectures





A- One stage architecture

B- Two stage architecture

(c) Presentation of detection architectures

Figure 15. Example of common architures used for plant phenotyping traits

Figure 16. Example of different detections modalities on wheat)

as illustrated in figure 17 alternatives are available for another type of data. For instance, in Neural Language Processing, recurring neural networks (RNN) were used to process text data. It includes as a basic bloc a gated recurring unit (GRU) instead of a convolutional layer. GRU can process a data point with the information of previous data points: predicting the next word or sentence or the next value in the case of temporal series. The NLP community also developed Transformers, networks which are based on the attention mechanism. The attention mechanism can learn the correlation between two variables for all inputs. The attention mechanism can be applied in an unsupervised manner with encoder-decoder architecture. It is then called self-attention. It has been critical to the performances of the DeepMind solutions to solve the go game and protein folding. It is interesting to notice that such a "Vision Transformer" can reach state-of-the-art performances without any convolutional layer, despite being more sensitive to shapes than texture compared to CNN networks [115]. It has already been extended for detection, and segmentation [114]. On top of the new neural network architecture, progress has been achieved for non-grid structures such as 3D meshes or protein structures usually represented as graphs. Graph Attention Networks [124] extend the attention mechanism by taking as input the nodes features and the adjacent matrix as input.

1.3.2 Training robust DL algorithm for plant phenotyping can be tricky

The bias problem

When DL algorithms are trained, it is assumed that both the training and test dataset are independent and identically distributed (IID hypothesis). Creating such a dataset requires capturing most of the ordinary conditions of variations of an object of interest. A consistent distribution is called "domain". Defining the domain is entirely open depending on which statistics we want to use to determine the distribution: is it a set of expert metadata? Or statistics of the extracted features? Determining the exact distribution of a set of images, which are high dimensional objects, is not as easy as describing the distribution of 2-D or 3-D vectors. Using metadata or expert information is the first step to describe a domain. For instance, a set of images of cars in "Los Angeles streets in summer" can form a domain defined by specific weather or a specific architecture. Training an algorithm on such a dataset and testing it in a set of images of cars from "Minnesota in winter" could results in a dramatic

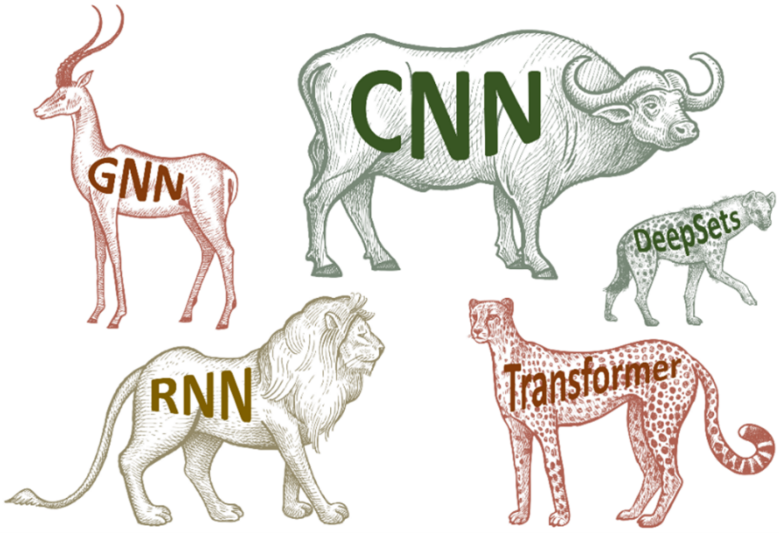


Figure 17. Illustration of the diversity of Deep Learning methods)

drop in performances, as the weather, the illumination, the architecture was never shown to the model previously. We call the "domain shift": the drop of performances on a new domain. It is then critical to have a dataset that covers the whole distribution of the "application domain" as illustrated in figure 18: the expected distribution of the data to predict. Some critical ethical issues were raised when using DL algorithms: Google's algorithm deployed for its photography application was mistaken "black people" for "gorilla" **??**. It was due to a wrong design of the training dataset that did not contain enough examples, and the evaluation was not robust enough.

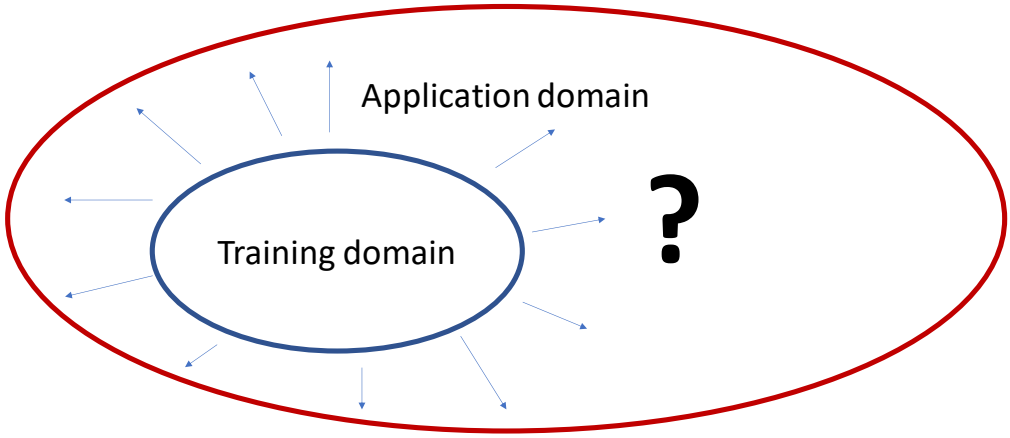


Figure 18. Illustration of the shift between application and train domain

It is crucial in plant phenotyping to control the IID hypothesis, mainly because we expect the error computed on the evaluation dataset to be the lowest possible and randomly spread. It is hence not acceptable to have a DL error that will vary with the domain. Many sources of variation exist in plant phenotyping. The first strong one is the difference between indoor and field conditions. Indoor images often correspond to individual plants, with controlled

illumination and a uniform background: the only difference between two images is then the expression of the genotype to the proposed environment and phenological stages. The difficulty is high in field conditions due to multiple, overlapping plants, variable lighting, and background. The environment influences more the characteristics of the acquired data. Genetics and environment for in-field phenotyping are then sources of variations. Acquisition conditions (flash use, RGB sensors quality and configuration) influence the image quality. Usually, data in phenotyping are acquired during an "acquisition session": a set of images that are acquired on the same experiment field, during a short time window with the same vector and sensors". Such acquisition sessions are made of highly similar images. Only the possible modalities within the experiment, and changes in illumination conditions, will bring additional variability to the genetic one. Acquisition sessions correspond, therefore, to very narrow distributions of images compared to the application domain, which generally corresponds to other acquisition sessions. The DL models are expected to be robust to the genetic, the environment and the sensor as described in figure 20. Training a DL model for plant phenotyping on only a few acquisition sessions does not ensure robustness. It does not prevent learning spurious correlation and uses features that are present only in the training domain. The goal of DL for plant phenotyping is to learn generic features that can extend to an "out-of-training domain" object.

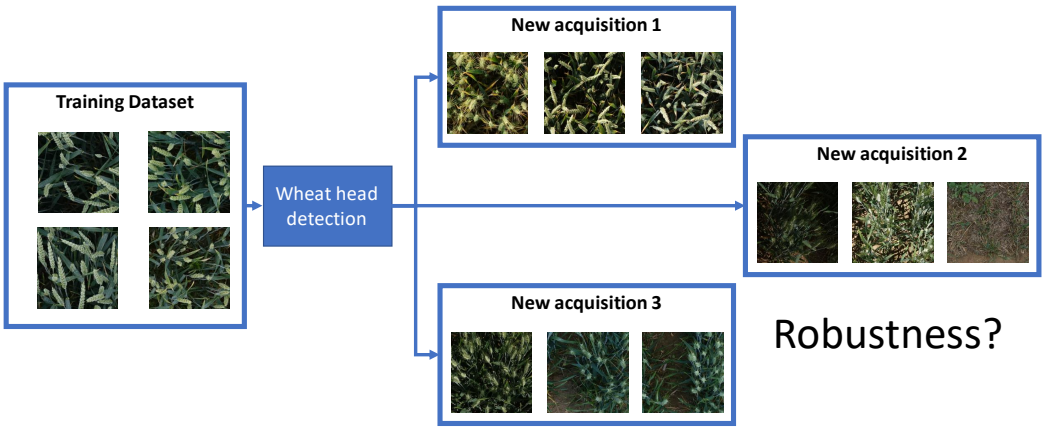


Figure 19. Illustration of the possible factor of variations

Building a large and diverse dataset

A solution to represent a data distribution covering all applications cases is to have a vast training dataset. A training dataset comprises data points, i.e. images, an audio file, text, and a corresponding target. For classification, the target is a "label" (a category); in semantic segmentation, it is a mask containing one or several labels per pixel; in object detection, it is a set of bounding boxes that enclose the aimed objects. Targets are created through a process called "annotation" or "labelling", where a human generate the target interactively. This process can be burdensome, depending on the interface used. The computer vision community relies on few large key datasets to benchmark their approach. As discussed in 1.3.1, ImageNet or MS COCO were key for DL development. In contrast, the plant phenotyping community suffers from a lack of such tools. Initiatives like the Leaf segmentation and counting datasets [125] exist but were limited to indoor phenotyping until today. Datasets can be found on platforms such as www.quantitative-plant.org or on more

general platforms such as Kaggle or Alcrowd that host several plant phenotyping datasets linked to actual or past competitions. The PlantVillage dataset (disease classification) [126] or PlantCLEF [127][128][129] are large datasets available on internet. Due to the diversity of traits to evaluate, there is a strong need to label even more data, and the current shared data is still not sufficient.

This issue can first be viewed as an engineering problem to solve. Several companies built labelling tools that fulfil all DL needs, including plant phenotyping. Such interfaces are progressing fast and have started to propose "smart tools" to accelerate the creation of manual targets. Figure 20 presents an advanced labeling platform (Datatorch). DEXTER (Deep Extreme Cut) [130] is an example of innovation for labelling, by allowing to create masks of a single instance by indicating only a few points on the frontiers, by leveraging deep learning features to find the contour of the object. Laboratories and institutes can develop additional tools to enhance the labelling experience: the CAPTE team has developed a tool that labels superpixels instead of single-pixel to create segmentation masks. In contrast with DL models primarily shared in open-source, labelling innovations start to be proprietary and less described; for example, the "smart tool" proposed by V7-Darwin can infer an object mask from a loose box and interact with the user, but the technology behind is private.

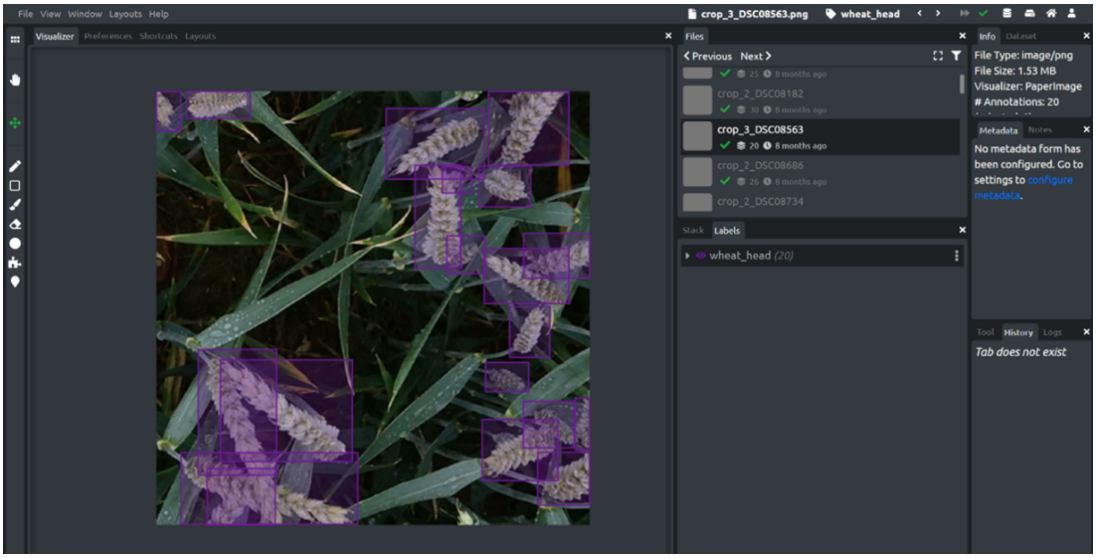


Figure 20. Screenshot of a Deep Learning annotation platform

The tedious nature of the labelling process leads to creative developments to accelerate the process. A direct way is to leverage the previously accumulated datasets to train the DL model and propose possible targets that must be validated or corrected by a human. These approaches have been tested in plant phenotyping by computing the outputs of a sorghum head detector on new images and letting a human correcting the output [131]. It is claimed to accelerate the process by a factor of 4. An alternative is to re-use "classic" computer vision algorithms to generate the possible targets. The SNORKEL [132] algorithm combines several heuristics to produce large training datasets for the DL algorithm. Recently, Liao et al. [133] proposed a method to understand the strength and weaknesses of human labellers, which have all different priors and exploit them for a more efficient process. Generating large datasets to label is burdensome for small research teams, but more and more companies propose to label images at a competitive price.

Data augmentation

Long training using the same data points can reduce results due to the overfitting phenomena. More diversity can be achieved easily with the use of data augmentation when loading the training data. Data augmentation is a process to generate a valid variation of

the training data. It can be easily implemented for straightforward strategies that change the image's geometry with rotation, padding or shearing, for instance or the photogrammetry with additional blur or change in the colour histogram. Simple perturbations such as Gaussian noise are also a form of Data Augmentation. The diversity introduced works better if the resulting images still belong to the application domain. For example, a random variation in the image's contrast should result in a new image where the objects of interest are still visible. Expert knowledge is then crucial to choose the proper data augmentation, though it can also be determined with automatic approaches such as AutoAugment [134]. Most of the usual Data Augmentations are included in python libraries such as albumentations [135] which can be integrated into DL frameworks such as Pytorch [136], or Keras [137].

The challenge of overcoming bias without very large datasets

Even with improvements in labelling, and possible outsourcing, it is desirable to constrain the learning process to get more robust networks. We can leverage three possible ideas: transductive transfer learning, invariant features learning or modelling the diversity.

- ***Transductive transfer learning*** is close to the idea of active learning but without human supervision. Its most simple form is pseudo-labelling[138], where we apply the model to the new domain. However, suppose the performance is still acceptable, the predictions can be used as labels, called "weak labels", and re-train the model jointly on the clean and weak labels, so the classifier is adapted to the aimed domain. It works when the source and the target domain are close and can help to boost the performance. However, while being an excellent solution to align the features from the source and target domain, it is not explicitly doing it.
- ***Domain adversarial model*** [139], [140] propose to append a second classifier that tries to predict the domain, but with a Gradient Reversal Layer (GRU) that invert the gradients to force the network not to discriminate the domain. The advantage of the method is that it does not require any additional labels on the task for the target domain and can be trained only on the source domain labels. Instead of using a classifier, DeepCORAL [141] proposes to add a term on the loss function that constrains the cross-correlation of the features network between the train and the source domain. These algorithms can be extended to n domains. The domain adaptation can also be accomplished during the training. The loss function is usually calculated without prior on the domain and is an "Empirical Risk Minimization". In contrast, the "Invariant Risk Minimization" (IRM) [142] has been proposed to force the classifier to generalize over the distributions of the train domains and is expected to work better on out-of-distribution objects. Other algorithms such as GroupDRO [143] propose to minimize the worst performance case to reach generalization.
- ***Generative Adversarial network***: The third idea is to learn the model the diversity of the dataset and generate more examples. A generative Adversarial Network learns to generate a large diversity of images. Models such as CycleGAN [144] can model the transformation required between images from two domains without any labels or paired images. It is an appealing solution to model a domain. However, using the output of a CycleGAN has not yet been demonstrated to solve the domain shift problem.

1.3.3 How to evaluate the robustness of DL models ?

Pitfalls of the current Deep Learning evaluation benchmark

DL community often works on one or a few datasets such as MNIST, CIFAR-10, CIFAR-100, CelebA, MS COCO, or ImageNet to benchmark the models. These datasets contain several limits: a study in 2021 [145] have identified an average of 3.4% of error in the labels. Some examples are shown in figure 21. After correcting the errors, the study demonstrates that

shallower models such as ResNet 34 perform better than their deeper counterpart, ResNet-50. Recht et al. [146] extended the ImageNet test set by proposing a similar yet different test set and observe an averaged 11-14% increase of performances. It is difficult in this context to conclude that a slight improvement (1-3%) on a benchmark dataset translates to an actual improvement on any other independent dataset. It explains why some "old" models are still popular despite the many new propositions published every year. The small number of metrics used to evaluate the model performances is also problematic. Their usefulness for benchmarking models is generally not discussed, but standard metrics are still widely used. For instance, the popular intersection over union (IoU) measures how well two bounding boxes match. However, when they do not overlap, the IoU is always equal to 0, whatever the distance between the two boxes. Rezatofighi [147] proposed a new metric to correct this weakness. Using a single metric is therefore not sufficient to thoroughly benchmark models. Further, most of the mentioned datasets propose a train, validation, and test dataset part of the same distribution and then respect the IID hypothesis. It is also valid for plant counting, as shown in our work [148]. Work on domain shift was conducted on shifts that are not likely to happen in real life: mixing drawings to images [149]; fake snow or flare, change of colour of MNIST digits [142].

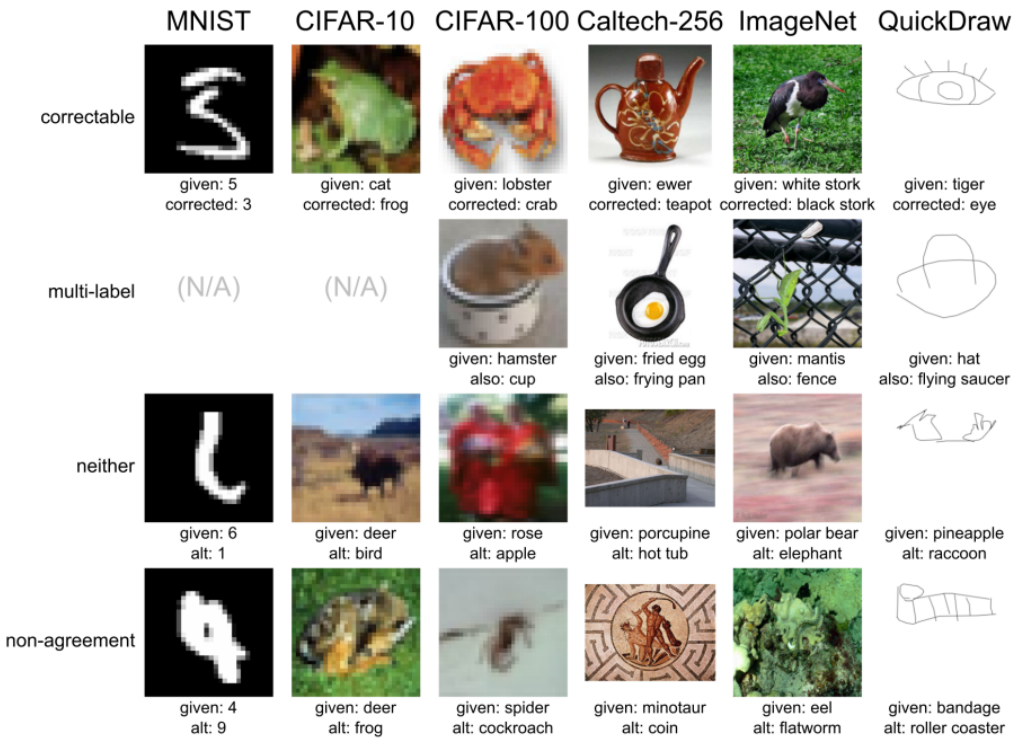


Figure 1: An example label error from each category (Sec. 5) for image datasets. The figure shows given labels, human-validated corrected labels, also the second label for multi-class data points, and CL-guessed alternatives. A browser for all label errors across all 10 datasets is available at https://labelerrors.com. Errors from text and audio datasets are also included on the website.

Figure 21. Visualization of label error in well-know datasets. reproduced from [145] with authorization of the author

Assessing the robustness of traits estimated with Deep Learning

Assessing the robustness of a trait estimated with DL is necessary to ensure its quality when deployed operationally. The dataset used for the evaluation needs to cover the range of environmental variability, the genotypic diversity, the development stages, the sensors,

and the different acquisition protocols used. Further, a part of the test dataset needs to be labelled the same way as the training one. However, because of the effort required for labelling, it is not always possible to represent the whole expected variability. Alternatives exist, such as labelling the test dataset with easier labels that cannot be used for training but still correlates with the measurement of the trait. The masks used for the segmentation task can be replaced by a set of labelled pixels, or the set of boxes used for detection can be replaced by a set of points on the objects of interest. Visual inspection is also helpful to find meaningful types of errors. Tools can be used to interpret DL algorithm outputs to understand how they took a decision. Visualizing intermediate outputs can also help to understand possible mistakes. Algorithms such as GradCAM [150] are used to interpret outputs of CNN, but new architectures such as transformer directly learn the correlation between pixels of an image. Most networks also output a confidence score with their predictions, which can indicate possible problems. However, high confidence scores are also observed on erroneous predictions when the model is poorly calibrated [151].

The robustness of the DL method can be evaluated by comparison with independent measurements of the trait. Traits measured destructively, such as the biomass or the chlorophyll content, can provide a solid reference. Manual measurements also produce useful references, although they can vary between field workers and may depend on their fatigue. A perfect match between the trait estimates and the independent measurement is therefore not always desired. However, the error between the DL estimates and the independent measurement needs to be as low as possible, especially the possible bias. For phenotyping applications, the error should not depend on the genotype, and this should be systematically checked, although it is rarely the case. Another difficulty arises from the difference in spatial support used for the DL input measurements, i.e. the image footprint and the independent measurement. For example, disease scoring is generally based on the inspection of the whole microplot, while the images used as input to the DL model cover only a sample of the microplot. Ideally, the spatial support should be the same for the DL trait estimates and their independent measurement.

Additional consistency tests can be conducted, even when no reference method is available. The broad-sense heritability [152], or reliability [153], can be used as a quality indicator by describing how repeatable the measurement is when exploiting the repetitions within a specifically designed experimental plan. The consistency between the trait estimated and another one, such as the final yield or critical performance characteristics such as lodging, disease resistance, or a particular yield component, can be a helpful indicator [45]. Finally, the quality of the possible association between a trait and markers of the genome, as found in a genome-wide association study (GWAS), can indicate consistency. However, these consistency checks are not sufficient since the estimated traits can be confounded with other highly heritable traits and possibly associated with some genome markers.

1.4 Objectives and organization of the study

The study aims to investigate possible strategies to train deep learning models that can be used operationally for high-throughput plant phenotyping experiments under field conditions. Special attention will be brought to the robustness of the estimates. For this purpose, the study relies on diverse and extended datasets collected over several partners and covers several traits and several species. The study is organized into four main parts, the three first ones corresponding to articles published, submitted, or to be submitted to international journals:

1. The first chapter discusses the robustness of DL method to identify plants from high-resolution RGB imagery against handcrafted methods that rely on expert knowledge. This corresponds to an article submitted to European Journal of Agronomy.
2. The second chapter focuses on the building of a very large datasets of wheat head

gathered from different institutions, and how it has been updated. It corresponds to two articles that have been published in Plant Phenomics.
3. The third chapter explains how this dataset has been leveraged to train robust wheat head detector thanks to the collaboration of competitors across the world for two successive challenges. The corresponding article is to be submitted to Gigasciences
4. We finally discuss our results in a final chapter that determine the perspective of Deep Learning for plants phenotyping. The chapter is based on an article published at ICML.

1.5 References

[1] *Food systems | ifpri : International food policy research institute*. [Online]. Available: https://www.ifpri.org/topic/food-systems.

[2] N. G. C. Change, *Global surface temperature | nasa global climate change*. [Online]. Available: https://climate.nasa.gov/vital-signs/global-temperature.

[3] *Meteo france - bilans climatiques*. [Online]. Available: http://www.meteofrance.fr/climat-passe-et-futur/bilans-climatiques.

[4] J. Hutchinson, C. Campbell, and R. Desjardins, "Some perspectives on carbon sequestration in agriculture," *Agricultural and forest meteorology*, vol. 142, no. 2-4, pp. 288–302, 2007.

[5] S. Asseng, F. Ewert, P. Martre, R. P. Rötter, D. B. Lobell, D. Cammarano, B. A. Kimball, M. J. Ottman, G. W. Wall, J. W. White, M. P. Reynolds, P. D. Alderman, P. V. V. Prasad, P. K. Aggarwal, J. Anothai, B. Basso, C. Biernath, A. J. Challinor, G. De Sanctis, J. Doltra, E. Fereres, M. Garcia-Vila, S. Gayler, G. Hoogenboom, L. A. Hunt, R. C. Izaurralde, M. Jabloun, C. D. Jones, K. C. Kersebaum, A.-K. Koehler, C. Müller, S. Naresh Kumar, C. Nendel, G. O'Leary, J. E. Olesen, T. Palosuo, E. Priesack, E. Eyshi Rezaei, A. C. Ruane, M. A. Semenov, I. Shcherbak, C. Stöckle, P. Stratonovitch, T. Streck, I. Supit, F. Tao, P. J. Thorburn, K. Waha, E. Wang, D. Wallach, J. Wolf, Z. Zhao, and Y. Zhu, "Rising temperatures reduce global wheat production," en, *Nature Climate Change*, vol. 5, no. 2, pp. 143–147, Feb. 2015, ISSN: 1758-678X, 1758-6798. DOI: 10.1038/nclimate2470.

[6] C. Zhao, B. Liu, S. Piao, X. Wang, D. B. Lobell, Y. Huang, M. Huang, Y. Yao, S. Bassu, P. Ciais, J.-L. Durand, J. Elliott, F. Ewert, I. A. Janssens, T. Li, E. Lin, Q. Liu, P. Martre, C. Müller, S. Peng, J. Peñuelas, A. C. Ruane, D. Wallach, T. Wang, D. Wu, Z. Liu, Y. Zhu, Z. Zhu, and S. Asseng, "Temperature increase reduces global yields of major crops in four independent estimates," en, *Proceedings of the National Academy of Sciences*, vol. 114, no. 35, pp. 9326–9331, Aug. 29, 2017, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1701762114.

[7] *Food and agriculture statistics*, en. [Online]. Available: http://www.fao.org/food-agriculture-statistics/en/.

[8] *Obesity and overweight*, en. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[9] A. M. Prentice, "The emerging epidemic of obesity in developing countries," en, *International Journal of Epidemiology*, vol. 35, no. 1, pp. 93–99, Feb. 1, 2006, ISSN: 1464-3685, 0300-5771. DOI: 10.1093/ije/dyi272.

[10] FAO and IFPRI, *Progress towards ending hunger and malnutrition: A cross-country cluster analysis*, en. Rome, Italy: FAO and IFPRI, 2020, ISBN: 978-92-5-132400-4. [Online]. Available: http://www.fao.org/documents/card/en/c/ca8593en.

[11] *World population*, en, May 23, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=World_populationI%5C&oldid=1024746492.

[12] W. Aktar, D. Sengupta, and A. Chowdhury, "Impact of pesticides use in agriculture: Their benefits and hazards," *Interdisciplinary toxicology*, vol. 2, no. 1, pp. 1–12, 2009.

[13] K. Norris, "Agriculture and biodiversity conservation: Opportunity knocks," *Conservation letters*, vol. 1, no. 1, pp. 2–11, 2008.

[14] B. J. Cardinale, J. E. Duffy, A. Gonzalez, D. U. Hooper, C. Perrings, P. Venail, A. Narwani, G. M. Mace, D. Tilman, D. A. Wardle, *et al.*, "Biodiversity loss and its impact on humanity," *Nature*, vol. 486, no. 7401, pp. 59–67, 2012.

[15] K. Brooks, "Food for thought on the gmo debate," *Nature Biotechnology*, vol. 19, no. 8, pp. 711–711, Aug. 1, 2001, ISSN: 1546-1696. DOI: 10.1038/90738.

[16] C. Marris, "Public views on gmos: Deconstructing the myths: Stakeholders in the gmo debate often describe public opinion as irrational. but do they really understand the public?" *EMBO reports*, vol. 2, no. 7, pp. 545–548, 2001.

[17] S. Hielscher, I. Pies, V. Valentinov, and L. Chatalova, "Rationalizing the gmo debate: The ordonomic approach to addressing agricultural myths," *International journal of environmental research and public health*, vol. 13, no. 5, p. 476, 2016.

[18] M. Araki and T. Ishii, "Towards social acceptance of plant breeding by genome editing," *Trends in plant science*, vol. 20, no. 3, pp. 145–149, 2015.

[19] D. F. Calderini and G. A. Slafer, "Changes in yield and yield stability in wheat during the 20th century," *Field Crops Research*, vol. 57, no. 3, pp. 335–347, 1998.

[20] L. Evans and R. Fischer, "Yield potential: Its definition, measurement, and significance," *Crop science*, vol. 39, no. 6, pp. 1544–1551, 1999.

[21] M. D. Purugganan and D. Q. Fuller, "The nature of selection during plant domestication," *Nature*, vol. 457, no. 7231, pp. 843–848, 2009.

[22] J. H. Peng, D. Sun, and E. Nevo, "Domestication evolution, genetics and genomics in wheat," *Molecular Breeding*, vol. 28, no. 3, pp. 281–301, 2011.

[23] R.-L. Wang, A. Stec, J. Hey, L. Lukens, and J. Doebley, "The limits of selection during maize domestication," *Nature*, vol. 398, no. 6724, pp. 236–239, 1999.

[24] M. J. Kovach, M. T. Sweeney, and S. R. McCouch, "New insights into the history of rice domestication," *TRENDS in Genetics*, vol. 23, no. 11, pp. 578–587, 2007.

[25] C. Li, A. Zhou, and T. Sang, "Rice domestication by reducing shattering," *science*, vol. 311, no. 5769, pp. 1936–1939, 2006.

[26] J. De Wet and J. Harlan, "The origin and domestication of sorghum bicolor," *Economic Botany*, vol. 25, no. 2, pp. 128–135, 1971.

[27] O. Smith, W. V. Nicholson, L. Kistler, E. Mace, A. Clapham, P. Rose, C. Stevens, R. Ware, S. Samavedam, G. Barker, *et al.*, "A domestication history of dynamic adaptation and genomic deterioration in sorghum," *Nature plants*, vol. 5, no. 4, pp. 369–379, 2019.

[28] F. Winchell, C. J. Stevens, C. Murphy, L. Champion, and D. Q. Fuller, "Evidence for sorghum domestication in fourth millennium bc eastern sudan: Spikelet morphology from ceramic impressions of the butana group," *Current Anthropology*, vol. 58, no. 5, pp. 673–683, 2017.

[29] Y. N. Harari, *Sapiens: A brief history of humankind*. Random House, 2014.

[30] R. A. Fisher, "Design of experiments," *Br Med J*, vol. 1, no. 3923, pp. 554–554, 1936.

[31] P. B. Hazell, C. Ramasamy, *et al.*, *The Green Revolution reconsidered: the impact of high-yielding rice varieties in South India.* Johns Hopkins University Press, 1991.

[32] P. L. Pingali, "Green revolution: Impacts, limits, and the path ahead," *Proceedings of the National Academy of Sciences*, vol. 109, no. 31, pp. 12 302–12 308, 2012.

[33] C. M. Hoffmann and C. Kenter, "Yield potential of sugar beet – have we hit the ceiling?" *Frontiers in Plant Science*, vol. 9, p. 289, 2018, ISSN: 1664-462X. DOI: 10.3389/fpls.2018.00289.

[34] *Snp genotyping*, en, Mar. 24, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=SNP_genotyping|%5C&oldid=1013952214.

[35] Y. Kawahara, M. de la Bastide, J. P. Hamilton, H. Kanamori, W. R. McCombie, S. Ouyang, D. C. Schwartz, T. Tanaka, J. Wu, S. Zhou, K. L. Childs, R. M. Davidson, H. Lin, L. Quesada-Ocampo, B. Vaillancourt, H. Sakai, S. S. Lee, J. Kim, H. Numa, T. Itoh, C. R. Buell, and T. Matsumoto, "Improvement of the oryza sativa nipponbare reference genome using next generation sequence and optical map data," *Rice*, vol. 6, no. 1, p. 4, Feb. 6, 2013, ISSN: 1939-8433. DOI: 10.1186/1939-8433-6-4.

[36] K. F. X. Mayer, R. Waugh, P. Langridge, T. J. Close, R. P. Wise, A. Graner, T. Matsumoto, K. Sato, A. Schulman, G. J. Muehlbauer, N. Stein, R. Ariyadasa, D. Schulte, N. Poursarebani, R. Zhou, B. Steuernagel, M. Mascher, U. Scholz, B. Shi, P. Langridge, K. Madishetty, J. T. Svensson, P. Bhat, M. Moscou, J. Resnik, T. J. Close, G. J. Muehlbauer, P. Hedley, H. Liu, J. Morris, R. Waugh, Z. Frenkel, A. Korol, H. Bergès, A. Graner, N. Stein, B. Steuernagel, U. Scholz, S. Taudien, M. Felder, M. Groth, M. Platzer, N. Stein, B. Steuernagel, U. Scholz, A. Himmelbach, S. Taudien, M. Felder, M. Platzer, S. Lonardi, D. Duma, M. Alpert, F. Cordero, M. Beccuti, G. Ciardo, Y. Ma, S. Wanamaker, T. J. Close, N. Stein, F. Cattonaro, V. Vendramin, S. Scalabrin, S. Radovic, R. Wing, D. Schulte, B. Steuernagel, M. Morgante, N. Stein, R. Waugh, T. Nussbaumer, H. Gundlach, M. Martis, R. Ariyadasa, N. Poursarebani, B. Steuernagel, U. Scholz, R. P. Wise, J. Poland, N. Stein, K. F. X. Mayer, M. Spannagl, M. Pfeifer, H. Gundlach, K. F. X. Mayer, H. Gundlach, C. Moisy, J. Tanskanen, S. Scalabrin, A. Zuccolo, V. Vendramin, M. Morgante, K. F. X. Mayer, A. Schulman, M. Pfeifer, M. Spannagl, P. Hedley, J. Morris, J. Russell, A. Druka, D. Marshall, M. Bayer, D. Swarbreck, D. Sampath, S. Ayling, M. Febrer, M. Caccamo, T. Matsumoto, T. Tanaka, K. Sato, R. P. Wise, T. J. Close, S. Wannamaker, G. J. Muehlbauer, N. Stein, K. F. X. Mayer, R. Waugh, B. Steuernagel, T. Schmutzer, M. Mascher, U. Scholz, S. Taudien, M. Platzer, K. Sato, D. Marshall, M. Bayer, R. Waugh, N. Stein, K. F. X. Mayer, R. Waugh, J. W. S. Brown, A. Schulman, P. Langridge, M. Platzer, G. B. Fincher, G. J. Muehlbauer, K. Sato, T. J. Close, R. P. Wise, N. Stein, T. I. B. G. S. Consortium, P. investigators, P. map construction, direct anchoring, G. sequencing, assembly, B. sequencing, assembly, B.-e. sequencing, I. of physical/genetic map, sequence resources, G. annotation, R. D. analysis, T. sequencing, analysis, Re-sequencing, diversity analysis, Writing, and editing of the manuscript, "A physical, genetic and functional sequence assembly of the barley genome," *Nature*, vol. 491, no. 7426, pp. 711–716, Nov. 1, 2012, ISSN: 1476-4687. DOI: 10.1038/nature11543.

[37] P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C.-T.

Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A.-P. Hsia, W. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J.-M. Chia, J.-M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddeloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. H. Nagel, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L. Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing, and R. K. Wilson, "The b73 maize genome: Complexity, diversity, and dynamics," *Science*, vol. 326, no. 5956, pp. 1112–1115, 2009, ISSN: 0036-8075. DOI: 10.1126/science.1178534.

[38] J. Schmutz, S. B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D. L. Hyten, Q. Song, J. J. Thelen, J. Cheng, *et al.*, "Genome sequence of the palaeopolyploid soybean," *nature*, vol. 463, no. 7278, pp. 178–183, 2010.

[39] T. I. W. G. S. C. (IWGSC), R. Appels, K. Eversole, N. Stein, C. Feuillet, B. Keller, J. Rogers, C. J. Pozniak, F. Choulet, A. Distelfeld, J. Poland, G. Ronen, A. G. Sharpe, O. Barad, K. Baruch, G. Keeble-Gagnère, M. Mascher, G. Ben-Zvi, A.-A. Josselin, A. Himmelbach, F. Balfourier, J. Gutierrez-Gonzalez, M. Hayden, C. Koh, G. Muehlbauer, R. K. Pasam, E. Paux, P. Rigault, J. Tibbits, V. Tiwari, M. Spannagl, D. Lang, H. Gundlach, G. Haberer, K. F. X. Mayer, D. Ormanbekova, V. Prade, H. Šimková, T. Wicker, D. Swarbreck, H. Rimbert, M. Felder, N. Guilhot, G. Kaithakottil, J. Keilwagen, P. Leroy, T. Lux, S. Twardziok, L. Venturini, A. Juhász, M. Abrouk, I. Fischer, C. Uauy, P. Borrill, R. H. Ramirez-Gonzalez, D. Arnaud, S. Chalabi, B. Chalhoub, A. Cory, R. Datla, M. W. Davey, J. Jacobs, S. J. Robinson, B. Steuernagel, F. van Ex, B. B. H. Wulff, M. Benhamed, A. Bendahmane, L. Concia, D. Latrasse, J. Bartoš, A. Bellec, H. Berges, J. Doležel, Z. Frenkel, B. Gill, A. Korol, T. Letellier, O.-A. Olsen, K. Singh, M. Valárik, E. van der Vossen, S. Vautrin, S. Weining, T. Fahima, V. Glikson, D. Raats, J. Číhalíková, H. Toegelová, J. Vrána, P. Sourdille, B. Darrier, D. Barabaschi, L. Cattivelli, P. Hernandez, S. Galvez, H. Budak, J. D. G. Jones, K. Witek, G. Yu, I. Small, J. Melonek, R. Zhou, T. Belova, K. Kanyuka, R. King, K. Nilsen, S. Walkowiak, R. Cuthbert, R. Knox, K. Wiebe, D. Xiang, A. Rohde, T. Golds, J. Čížková, B. A. Akpinar, S. Biyiklioglu, L. Gao, A. N'Daiye, M. Kubaláková, J. Šafář, F. Alfama, A.-F. Adam-Blondon, R. Flores, C. Guerche, M. Loaec, H. Quesneville, J. Condie, J. Ens, R. Maclachlan, Y. Tan, A. Alberti, J.-M. Aury, V. Barbe, A. Couloux, C. Cruaud, K. Labadie, S. Mangenot, P. Wincker, G. Kaur, M. Luo, S. Sehgal, P. Chhuneja, O. P. Gupta, S. Jindal, P. Kaur, P. Malik, P. Sharma, B. Yadav, N. K. Singh, J. P. Khurana, C. Chaudhary, P. Khurana, V. Kumar, A. Mahato, S. Mathur, A. Sevanthi, N. Sharma, R. S. Tomar, K. Holušová, O. Plíhal, M. D. Clark, D. Heavens, G. Kettleborough, J. Wright, B. Balcárková, Y. Hu, E. Salina, N. Ravin, K. Skryabin, A. Beletsky, V. Kadnikov, A. Mardanov, M. Nesterov, A. Rakitin, E. Sergeeva, H. Handa, H. Kanamori, S. Katagiri, F. Kobayashi, S. Nasuda, T. Tanaka, J. Wu, F. Cattonaro, M. Jiumeng, K. Kugler, M. Pfeifer, S. Sandve, X. Xun, B. Zhan, J. Batley, P. E. Bayer, D. Edwards, S. Hayashi, Z. Tulpová, P. Visendi, L. Cui, X. Du, K. Feng, X. Nie, W. Tong, and L. Wang, "Shifting the limits in wheat research and breeding using a fully annotated reference genome," en, *Science*, vol. 361, no. 6403, eaar7191, Aug. 17, 2018, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aar7191.

[40] B. Mamta and M. V. Rajam, "Rnai technology: A new platform for crop pest control," eng, *Physiology and molecular biology of plants : an international journal of functional plant biology*, vol. 23, no. 3, pp. 487–501, Jul. 2017, ISSN: 0971-5894. DOI: 10.1007/s12298-017-0443-x.

[41] K. Massel, I. Godwin, and L. Hickey, "Tunable crops are just a spray away," *Nature Plants*, vol. 7, no. 2, pp. 102–103, Feb. 1, 2021, ISSN: 2055-0278. DOI: 10.1038/s41477-021-00849-6.

[42] F. A. Ran, P. D. Hsu, J. Wright, V. Agarwala, D. A. Scott, and F. Zhang, "Genome engineering using the crispr-cas9 system," en, *Nature Protocols*, vol. 8, no. 11, pp. 2281–2308, Nov. 2013, ISSN: 1750-2799. DOI: 10.1038/nprot.2013.143.

[43] S. Ahmad, X. Wei, Z. Sheng, P. Hu, and S. Tang, "CRISPR/Cas9 for development of disease resistance in plants: recent progress, limitations and future prospects," *Briefings in Functional Genomics*, vol. 19, no. 1, pp. 26–39, Jan. 2020, ISSN: 2041-2657. DOI: 10.1093/bfgp/elz041. eprint: https://academic.oup.com/bfg/article-pdf/19/1/26/32524954/elz041.pdf. [Online]. Available: https://doi.org/10.1093/bfgp/elz041.

[44] P. Martres, B. Quilot-Turion, D. Luquet, M. MEMMAH, K. Chenu, and P. Debaeke, "Model-assisted phenotyping and ideotype design," in *Crop Physiology*, Elsevier, 2015, np. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01315577.

[45] M. Reynolds, S. Chapman, L. Crespo-Herrera, G. Molero, S. Mondal, D. N. Pequeno, F. Pinto, F. J. Pinera-Chavez, J. Poland, C. Rivera-Amado, C. Saint Pierre, and S. Sukumaran, "Breeder friendly phenotyping," en, *Plant Science*, vol. 295, p. 110 396, Jun. 2020, ISSN: 01689452. DOI: 10.1016/j.plantsci.2019.110396.

[46] C. M. Caruso, C. M. Mason, and J. S. Medeiros, "The evolution of functional traits in plants: Is the giant still sleeping?" *International Journal of Plant Sciences*, vol. 181, no. 1, pp. 1–8, Jan. 1, 2020, ISSN: 1058-5893. DOI: 10.1086/707141.

[47] J. Blancon, D. Dutartre, M.-H. Tixier, M. Weiss, A. Comar, S. Praud, and F. Baret, "A high-throughput model-assisted method for phenotyping maize green leaf area index dynamics using unmanned aerial vehicle imagery," *Frontiers in Plant Science*, vol. 10, p. 685, 2019, ISSN: 1664-462X. DOI: 10.3389/fpls.2019.00685.

[48] F. Baret, "Contribution au suivi radiométrique de cultures de céréales," Ph.D. dissertation, 1986.

[49] J. Lizaso, W. Batchelor, and M. Westgate, "A leaf area model to simulate cultivar-specific expansion and senescence of maize leaves," *Field crops research*, vol. 80, no. 1, pp. 1–17, 2003.

[50] S. Liu, P. Martre, S. Buis, M. Abichou, B. Andrieu, and F. Baret, "Estimation of plant and canopy architectural traits using the digital plant phenotyping platform," en, *Plant Physiology*, vol. 181, no. 3, pp. 881–890, Nov. 2019, ISSN: 0032-0889, 1532-2548. DOI: 10.1104/pp.19.00554.

[51] B. Soenen, X. BRIS, M. Laberdesque, J. Cohan, F. Laurent, A. Bouthier, D. Gouache, and C. Garcia, ""chn", a crop model to jointly manage water and nitrogen on winter wheat," *Stress*, vol. 25, no. 30, 2016.

[52] G. L. Hammer, E. van Oosterom, G. McLean, S. C. Chapman, I. Broad, P. Harland, and R. C. Muchow, "Adapting apsim to model the physiology and genetics of complex adaptive traits in field crops," en, *Journal of Experimental Botany*, vol. 61, no. 8, pp. 2185–2202, May 2010, ISSN: 1460-2431, 0022-0957. DOI: 10.1093/jxb/erq095.

[53] D. Holzworth, N. Huth, J. Fainges, H. Brown, E. Zurcher, R. Cichota, S. Verrall, N. Herrmann, B. Zheng, and V. Snow, "Apsim next generation: Overcoming challenges in modernising a farming systems model," en, *Environmental Modelling I& Software*, vol. 103, pp. 43–51, May 2018, ISSN: 13648152. DOI: 10.1016/j.envsoft.2018.02.002.

[54] B. A. Keating, P. S. Carberry, G. L. Hammer, M. E. Probert, M. J. Robertson, D. Holzworth, N. I. Huth, J. N. Hargreaves, H. Meinke, Z. Hochman, *et al.*, "An overview of apsim, a model designed for farming systems simulation," *European journal of agronomy*, vol. 18, no. 3-4, pp. 267–288, 2003.

[55] N. Brisson, C. Gary, E. Justes, R. Roche, B. Mary, D. Ripoche, D. Zimmer, J. Sierra, P. Bertuzzi, P. Burger, *et al.*, "An overview of the crop model stics," *European Journal of agronomy*, vol. 18, no. 3-4, pp. 309–332, 2003.

[56] N. Brisson, B. Mary, D. Ripoche, M.-H. Jeuffroy, F. Ruget, B. Nicoullaud, P. Gate, F. Devienne-Barret, R. Antonioletti, C. Dürr, *et al.*, "Stics: A generic model for the simulation of crops and their water and nitrogen balances. i. theory and parameterization applied to wheat and corn," *Agronomie*, vol. 18, no. 5-6, pp. 311–346, 1998.

[57] N. Brisson, M. Launay, B. Mary, and N. Beaudoin, *Conceptual basis, formalisations and parameterization of the STICS crop model*. Editions Quae, 2009.

[58] C. M. Donald, "The breeding of crop ideotypes," en, *Euphytica*, vol. 17, no. 3, pp. 385–403, Dec. 1, 1968, ISSN: 1573-5060. DOI: 10.1007/BF00056241.

[59] J. Roy, F. Rineau, H. J. D. Boeck, I. Nijs, T. Pütz, S. Abiven, J. A. Arnone, C. V. M. Barton, N. Beenaerts, N. Brüggemann, M. Dainese, T. Domisch, N. Eisenhauer, S. Garré, A. Gebler, A. Ghirardo, R. L. Jasoni, G. Kowalchuk, D. Landais, S. H. Larsen, V. Leemans, J.-F. L. Galliard, B. Longdoz, F. Massol, T. N. Mikkelsen, G. Niedrist, C. Piel, O. Ravel, J. Sauze, A. Schmidt, J.-P. Schnitzler, L. H. Teixeira, M. G. Tjoelker, W. W. Weisser, B. Winkler, and A. Milcu, "Ecotrons: Powerful and versatile ecosystem analysers for ecology, agronomy and environmental science," en, *Global Change Biology*, vol. 27, no. 7, pp. 1387–1407, Apr. 1, 2021, ISSN: 1365-2486.

[60] J.-C. Deswarte, K. Beauchene, G. Arjaure, S. Jezequel, G. Meloux, Y. Flodrops, J. Landrieaux, A. Bouthier, S. Thomas, B. De Solan, *et al.*, "Platform development for drought tolerance evaluation of wheat in france," *Procedia Environmental Sciences*, vol. 29, pp. 93–94, 2015.

[61] N. Kirchgessner, F. Liebisch, K. Yu, J. Pfeifer, M. Friedli, A. Hund, and A. Walter, "The eth field phenotyping platform fip: A cable-suspended multi-sensor system," *Functional Plant Biology*, vol. 44, no. 1, pp. 154–168, 2016.

[62] N. Virlet, K. Sabermanesh, P. Sadeghi-Tehran, and M. J. Hawkesford, "Field scanalyzer: An automated robotic field phenotyping platform for detailed crop monitoring," *Functional Plant Biology*, vol. 44, no. 1, pp. 143–153, 2016.

[63] P. Andrade-Sanchez, M. A. Gore, J. T. Heun, K. R. Thorp, A. E. Carmo-Silva, A. N. French, M. E. Salvucci, and J. W. White, "Development and evaluation of a field-based high-throughput phenotyping platform," *Functional Plant Biology*, vol. 41, no. 1, pp. 68–79, 2013.

[64] T. Duan, B. Zheng, W. Guo, S. Ninomiya, Y. Guo, S. C. Chapman, T. Duan, B. Zheng, W. Guo, S. Ninomiya, Y. Guo, and S. C. Chapman, "Comparison of ground cover estimates from experiment plots in cotton, sorghum and sugarcane based on images and ortho-mosaics captured by uav," en, *Functional Plant Biology*, vol. 44, no. 1, pp. 169–183, Nov. 24, 2016, ISSN: 1445-4416, 1445-4416. DOI: 10.1071/FP16123.

[65] F. Gnädinger and U. Schmidhalter, "Digital counts of maize plants by unmanned aerial vehicles (uavs)," en, *Remote Sensing*, vol. 9, no. 6, p. 544, Jun. 2017. DOI: 10.3390/rs9060544.

[66] S. Liu, F. Baret, D. Allard, X. Jin, B. Andrieu, P. Burger, M. Hemmerlé, and A. Comar, "A method to estimate plant density and plant spacing heterogeneity: Application to wheat crops," *Plant methods*, vol. 13, no. 1, p. 38, 2017.

[67] S. Madec, F. Baret, B. de Solan, S. Thomas, D. Dutartre, S. Jezequel, M. Hemmerlé, G. Colombeau, and A. Comar, "High-throughput phenotyping of plant height: Comparing unmanned aerial vehicles and ground lidar estimates," en, *Frontiers in Plant Science*, vol. 8, p. 2002, Nov. 27, 2017, ISSN: 1664-462X. DOI: 10.3389/fpls.2017.02002.

[68] L. Mrisho, N. Mbilinyi, M. Ndalahwa, A. Ramcharan, A. Kehs, P. McCloskey, H. Murithi, D. Hughes, and J. Legg, "Evaluating the accuracy of a smartphone-based artificial intelligence system, plantvillage nuru, in diagnosing of the viral diseases of cassava," *bioRxiv*, 2020.

[69] K. Velumani, S. Madec, B. de Solan, R. Lopez-Lozano, J. Gillet, J. Labrosse, S. Jezequel, A. Comar, and F. Baret, "An automatic method based on daily in situ images and deep learning to date wheat heading stage," en, *Field Crops Research*, vol. 252, p. 107 793, Jul. 2020, ISSN: 03784290. DOI: 10.1016/j.fcr.2020.107793.

[70] M. Mishra, P. Choudhury, and B. Pati, "Modified ride-nn optimizer for the iot based plant disease detection," en, *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 1, pp. 691–703, Jan. 1, 2021, ISSN: 1868-5145. DOI: 10.1007/s12652-020-02051-6.

[71] A. Bannari, D. Morin, F. Bonn, and A. R. Huete, "A review of vegetation indices," *Remote Sensing Reviews*, vol. 13, no. 1-2, pp. 95–120, Aug. 1, 1995, ISSN: 0275-7257. DOI: 10.1080/02757259509532298.

[72] J. Xue and B. Su, "Significant remote sensing vegetation indices: A review of developments and applications," en, *Journal of Sensors*, vol. 2017, e1353691, May 23, 2017, ISSN: 1687-725X. DOI: 10.1155/2017/1353691.

[73] S. Jacquemoud, W. Verhoef, F. Baret, C. Bacour, P. J. Zarco-Tejada, G. P. Asner, C. François, and S. L. Ustin, "Prospect+sail models: A review of use for vegetation characterization," en, *Remote Sensing of Environment*, vol. 113, S56–S66, Sep. 2009, ISSN: 00344257. DOI: 10.1016/j.rse.2008.01.026.

[74] S. Jacquemoud and F. Baret, "Prospect: A model of leaf optical properties spectra," en, *Remote Sensing of Environment*, vol. 34, no. 2, pp. 75–91, Nov. 1, 1990, ISSN: 0034-4257. DOI: 10.1016/0034-4257(90)90100-Z.

[75] W. Verhoef, "Light scattering by leaf layers with application to canopy reflectance modeling: The sail model," en, *Remote Sensing of Environment*, vol. 16, no. 2, pp. 125–141, Oct. 1, 1984, ISSN: 0034-4257. DOI: 10.1016/0034-4257(84)90057-9.

[76] J. R. Ubbens and I. Stavness, "Deep plant phenomics: A deep learning platform for complex plant phenotyping tasks," en, *Frontiers in Plant Science*, vol. 8, p. 1190, Jul. 7, 2017, ISSN: 1664-462X. DOI: 10.3389/fpls.2017.01190.

[77] S. Dandrifosse, A. Bouvry, V. Leemans, B. Dumont, and B. Mercatoris, "Imaging wheat canopy through stereo vision: Overcoming the challenges of the laboratory to field transition for morphological features extraction," English, *Frontiers in Plant Science*, vol. 11, 2020, ISSN: 1664-462X. DOI: 10.3389/fpls.2020.00096. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpls.2020.00096/full#h4.

[78] S. Paulus, "Measuring crops in 3d: Using geometry for plant phenotyping," *Plant Methods*, vol. 15, no. 1, p. 103, Sep. 3, 2019, ISSN: 1746-4811. DOI: 10.1186/s13007-019-0490-0.

[79] A. Bucksch, J. Burridge, L. M. York, A. Das, E. Nord, J. S. Weitz, and J. P. Lynch, "Image-based high-throughput field phenotyping of crop roots," *Plant Physiology*, vol. 166, no. 2, pp. 470–486, Oct. 1, 2014, ISSN: 0032-0889. DOI: 10.1104/pp.114.243519.

[80] D. Kuznichov, A. Zvirin, Y. Honen, and R. Kimmel, "Data augmentation for leaf segmentation and counting tasks in rosette plants," en, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA: IEEE, Jun. 2019, pp. 2580–2589, ISBN: 978-1-72812-506-0. DOI: 10.1109/CVPRW.2019.00314. [Online]. Available: https://ieeexplore.ieee.org/document/9025429/.

[81] P. Sadeghi-Tehran, N. Virlet, K. Sabermanesh, and M. J. Hawkesford, "Multi-feature machine learning model for automatic segmentation of green fractional vegetation cover for high-throughput field phenotyping," en, *Plant Methods*, vol. 13, no. 1, p. 103, Dec. 2017, ISSN: 1746-4811. DOI: 10.1186/s13007-017-0253-8.

[82] Z. Fan, J. Lu, M. Gong, H. Xie, and E. D. Goodman, "Automatic tobacco plant detection in uav images via deep neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 876–887, Mar. 2018, ISSN: 2151-1535. DOI: 10.1109/JSTARS.2018.2793849.

[83] 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," en, *Agricultural and Forest Meteorology*, vol. 264, pp. 225–234, Jan. 15, 2019, ISSN: 0168-1923. DOI: 10.1016/j.agrformet.2018.10.013.

[84] A. Fuentes, S. Yoon, S. C. Kim, and D. S. Park, "A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition," en, *Sensors*, vol. 17, no. 9, p. 2022, Sep. 2017. DOI: 10.3390/s17092022.

[85] Y. Toda and F. Okura, "How convolutional neural networks diagnose plant disease," en, *Plant Phenomics*, vol. 2019, pp. 1–14, Mar. 26, 2019, ISSN: 2643-6515. DOI: 10.34133/2019/9237136.

[86] N. Shakoor, S. Lee, and T. C. Mockler, "High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field," en, *Current Opinion in Plant Biology*, vol. 38, pp. 184–192, Aug. 2017, ISSN: 13695266. DOI: 10.1016/j.pbi.2017.05.006.

[87] S. Aich, A. Josuttes, I. Ovsyannikov, K. Strueby, I. Ahmed, H. S. Duddu, C. Pozniak, S. Shirtliffe, and I. Stavness, "Deepwheat: Estimating phenotypic traits from crop images with deep learning," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Mar. 2018, pp. 323–332. DOI: 10.1109/WACV.2018.00042.

[88] M. F. Dreccer, G. Molero, C. Rivera-Amado, C. John-Bejai, and Z. Wilson, "Yielding to the image: How phenotyping reproductive growth can assist crop improvement and production," en, *Plant Science*, vol. 282, pp. 73–82, May 2019, ISSN: 01689452. DOI: 10.1016/j.plantsci.2018.06.008.

[89] M. P. Pound, J. A. Atkinson, A. J. Townsend, M. H. Wilson, M. Griffiths, A. S. Jackson, A. Bulat, G. Tzimiropoulos, D. M. Wells, E. H. Murchie, T. P. Pridmore, and A. P. French, "Deep machine learning provides state-of-the-art performance in image-based plant phenotyping," en, *GigaScience*, vol. 6, no. 10, Oct. 1, 2017, ISSN: 2047-217X. DOI: 10.1093/gigascience/gix083. [Online]. Available: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/gix083/4091592.

[90] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, May 2013, pp. 8599–8603. DOI: 10.1109/ICASSP.2013.6639344.

[91] Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, and M. S. Nasrin, "The history began from alexnet: A comprehensive survey on deep learning approaches," en, p. 39,

[92] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size," *arXiv:1602.07360 [cs]*, Nov. 4, 2016. [Online]. Available: http://arxiv.org/abs/1602.07360.

[93] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," 2017.

[94] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," en, *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 24, 2017, ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3065386.

[95] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2015, pp. 91–99.

[96] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[97] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2017, pp. 2961–2969.

[98] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," Springer, 2015, pp. 234–241.

[99] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473 [cs, stat]*, May 19, 2016. [Online]. Available: http://arxiv.org/abs/1409.0473.

[100] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958, ISSN: 1939-1471(Electronic),0033-295X(Print). DOI: 10.1037/h0042519.

[101] ——, "Analytic techniques for the study of neural nets," *IEEE Transactions on Applications and Industry*, vol. 83, no. 74, pp. 285–292, Sep. 1964, ISSN: 2379-6782. DOI: 10.1109/TAI.1964.5407758.

[102] *Backpropagation applied to handwritten zip code recognition | neural computation | mit press*. [Online]. Available: https://direct.mit.edu/neco/article-abstract/1/4/541/5515/Backpropagation-Applied-to-Handwritten-Zip-Code?redirectedFrom=fulltext.

[103] Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Müller, E. Säckinger, P. Simard, and V. Vapnik, "Comparison of learning algorithms for handwritten digit recognition," 1995, pp. 53–60.

[104] *Learning representations by back-propagating errors | nature*. [Online]. Available: https://www.nature.com/articles/323533a0.

[105] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," Ieee, 2009, pp. 248–255.

[106] T. Mensink, G. Csurka, F. Perronnin, J. Sánchez, and J. Verbeek, "Lear and xrce's participation to visual concept detection task - imageclef 2010," en, ImageCLEF - Workshop Cross Language Image Retrieval, Sep. 20, 2010, p. 48. [Online]. Available: https://hal.inria.fr/inria-00548633.

[107] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," en, *arXiv:1506.07503 [cs, stat]*, Jun. 24, 2015. [Online]. Available: http://arxiv.org/abs/1506.07503.

[108] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," en, *Science*, vol. 362, no. 6419, pp. 1140–1144, Dec. 7, 2018, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aar6404.

[109] *Improved protein structure prediction using potentials from deep learning | nature*. [Online]. Available: https://www.nature.com/articles/s41586-019-1923-7.

[110] R. Kleffner, J. Flatten, A. Leaver-Fay, D. Baker, J. B. Siegel, F. Khatib, and S. Cooper, "Foldit standalone: A video game-derived protein structure manipulation interface using rosetta," *Bioinformatics*, vol. 33, no. 17, pp. 2765–2767, 2017.

[111] J. Sanders and E. Kandrot, *CUDA by example: an introduction to general-purpose GPU programming*. Addison-Wesley Professional, 2010.

[112] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," Springer, 2014, pp. 740–755.

[113] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," *arXiv preprint arXiv:0809.3083*, 2008.

[114] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556 [cs]*, Apr. 10, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556.

[115] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*. 2015.

[116] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv:1905.11946 [cs, stat]*, Sep. 11, 2020. [Online]. Available: http://arxiv.org/abs/1905.11946.

[117] X. He, K. Zhao, and X. Chu, "Automl: A survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, p. 106 622, Jan. 2021, ISSN: 09507051. DOI: 10.1016/j.knosys.2020.106622.

[118] Y. You, Z. Zhang, C.-J. Hsieh, J. Demmel, and K. Keutzer, "Imagenet training in minutes," *arXiv:1709.05011 [cs]*, Jan. 31, 2018. [Online]. Available: http://arxiv.org/abs/1709.05011.

[119] *Now anyone can train imagenet in 18 minutes · fast.ai*. [Online]. Available: https://www.fast.ai/2018/08/10/fastai-diu-imagenet/.

[120] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv:1606.00915 [cs]*, May 11, 2017. [Online]. Available: http://arxiv.org/abs/1606.00915.

[121] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," Springer, 2016, pp. 21–37.

[122] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *arXiv:1506.02640 [cs]*, May 9, 2016. [Online]. Available: http://arxiv.org/abs/1506.02640.

[123] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861 [cs]*, Apr. 16, 2017. [Online]. Available: http://arxiv.org/abs/1704.04861.

[124] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," *arXiv:1710.10903 [cs, stat]*, Feb. 4, 2018. [Online]. Available: http://arxiv.org/abs/1710.10903.

[125] M. Minervini, A. Fischbach, H. Scharr, and S. A. Tsaftaris, "Finely-grained annotated datasets for image-based plant phenotyping," en, *Pattern Recognition Letters*, vol. 81, pp. 80–89, Oct. 2016, ISSN: 01678655. DOI: 10.1016/j.patrec.2015.10.013.

[126] D. P. Hughes and M. Salathé, "An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing," *CoRR*, vol. abs/1511.08060, 2015. [Online]. Available: http://arxiv.org/abs/1511.08060.

[127] A. Joly, H. Goëau, C. Botella, H. Glotin, P. Bonnet, W.-P. Vellinga, R. Planqué, and H. Müller, "Overview of lifeclef 2018: A large-scale evaluation of species identification and recommendation algorithms in the era of ai," en, in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, and N. Ferro, Eds., vol. 11018, Cham: Springer International Publishing, 2018, pp. 247–266, ISBN: 978-3-319-98931-0. [Online]. Available: http://link.springer.com/10.1007/978-3-319-98932-7_24.

[128] H. Goëau, P. Bonnet, and A. Joly, "Overview of lifeclef plant identification task 2020," 2020.

[129] ——, "Overview of lifeclef plant identification task 2019: Diving into data deficient tropical countries," vol. 2380, CEUR, 2019, pp. 1–13.

[130] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. V. Gool, "Deep extreme cut: From extreme points to object segmentation," 2018.

[131] S. Ghosal, B. Zheng, S. C. Chapman, A. B. Potgieter, D. R. Jordan, X. Wang, A. K. Singh, A. Singh, M. Hirafuji, S. Ninomiya, *et al.*, "A weakly supervised deep learning framework for sorghum head detection and counting," *Plant Phenomics*, vol. 2019, p. 1525874, 2019.

[132] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," vol. 11, NIH Public Access, 2017, p. 269.

[133] Y.-H. Liao, A. Kar, and S. Fidler, "Towards good practices for efficiently annotating large-scale image classification datasets," *arXiv:2104.12690 [cs]*, Apr. 26, 2021. [Online]. Available: http://arxiv.org/abs/2104.12690.

[134] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv:1805.09501 [cs, stat]*, Apr. 11, 2019. [Online]. Available: http://arxiv.org/abs/1805.09501.

[135] A. Buslaev, A. Parinov, E. Khvedchenya, I. V. I., and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *ArXiv e-prints*, 2018.

[136] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, d. F. Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[137] F. Chollet *et al.*, *Keras*, 2015. [Online]. Available: https://github.com/fchollet/keras.

[138] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," vol. 3, 2013.

[139] T. W. Ayalew, J. R. Ubbens, and I. Stavness, "Unsupervised domain adaptation for plant organ counting," en, p. 17,

[140] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv:1409.7495 [cs, stat]*, Feb. 27, 2015. [Online]. Available: http://arxiv.org/abs/1409.7495.

[141] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," *arXiv:1607.01719 [cs]*, Jul. 6, 2016. [Online]. Available: http://arxiv.org/abs/1607.01719.

[142] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv:1907.02893 [cs, stat]*, Mar. 27, 2020. [Online]. Available: http://arxiv.org/abs/1907.02893.

[143] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," en, *arXiv:1911.08731 [cs, stat]*, Apr. 2, 2020. [Online]. Available: http://arxiv.org/abs/1911.08731.

[144] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv:1703.10593 [cs]*, Aug. 24, 2020. [Online]. Available: http://arxiv.org/abs/1703.10593.

[145] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," *arXiv:2103.14749 [cs, stat]*, Apr. 8, 2021. [Online]. Available: http://arxiv.org/abs/2103.14749.

[146] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" *arXiv:1902.10811 [cs, stat]*, Jun. 12, 2019. [Online]. Available: http://arxiv.org/abs/1902.10811.

[147] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," *arXiv:1902.09630 [cs]*, Apr. 14, 2019. [Online]. Available: http://arxiv.org/abs/1902.09630.

[148] E. David, G. Daubige, F. Joudelat, P. Burger, A. Comar, B. De solan, and F. Baret, "Plant detection and counting from high-resolution rgb images acquired from uavs: Comparison between deep-learning and handcrafted methods with application to maize, sugar beet, and sunflower crops," *bioRxiv*, 2021. DOI: 10.1101/2021.04.27.441631. [Online]. Available: https://www.biorxiv.org/content/early/2021/04/28/2021.04.27.441631.

[149] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," en, 2017 IEEE International Conference on Computer Vision (ICCV), Venice: IEEE, Oct. 2017, pp. 5543–5551, ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.2017.591. [Online]. Available: http://ieeexplore.ieee.org/document/8237853/.

[150] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020, ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-019-01228-7.

[151] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *arXiv:1706.04599 [cs]*, Aug. 3, 2017. [Online]. Available: http://arxiv.org/abs/1706.04599.

[152] H.-P. Piepho and J. Möhring, "Computing heritability and selection response from unbalanced plant breeding trials," *Genetics*, vol. 177, no. 3, pp. 1881–1888, Nov. 2007, ISSN: 0016-6731. DOI: 10.1534/genetics.107.074229.

[153] R. Bernardo, "Reinventing quantitative genetics for plant breeding: Something old, something new, something borrowed, something blue," en, *Heredity*, vol. 125, no. 6, pp. 375–385, Dec. 2020, ISSN: 0018-067X, 1365-2540. DOI: 10.1038/s41437-020-0312-1.

2 Evaluation of the robustness of handcrafted and deep learning methods for plant density estimation

2.1 Foreword

Before the emergence of DL for phenotyping applications, some traits were already operationally extracted from high-resolution RGB imagery using computer vision classical approaches. Such "handcrafted" methods are based on expert knowledge on the problem from which features are identified and used to estimate the trait. The explicit use of well-verified assumptions can help scale to unseen conditions and, therefore, obtain more robust performances than DL methods that can be sensitive to domain shift. These handcrafted methods can also be used to post-process DL results and correct possible inappropriate ones. In this first chapter, we propose to compare a handcrafted approach with a DL approach for a plant detection problem. Plant detection is the mandatory first step before accessing higher-order traits such as plant density and plant characteristics. As reviewed in the introduction, we will focus on UAV observations that are a very efficient way to phenotype extensive experiments. It is very popular among breeders. However, the spatial resolution can be limited in case of too high altitude flights. We will consider in this chapter three species with relatively big leaves and sown with relatively even distances between plants: maize, sugar beet and sunflower.

Assessing the robustness of an approach is different from demonstrating the performance of new detection methods. While demonstrating the effectiveness of a new approach on a small dataset, evaluating its robustness needs to explore various possible conditions of acquisitions. In the introduction of this chapter, we demonstrate that most studies on plant counting or detection operate on a limited set of conditions, often with the validation achieved over the dataset used for the training of the DL model. A diverse dataset composed of UAV images from 27 flights (18 locations) covering three crops at different development stages has been compiled to compare a handcrafted approach with a DL approach. We further propose to combine both approaches into a hybrid one to beneficiate from their respective advantages.

2.2 Plant detection and counting from high-resolution RGB images acquired from UAVs

The following manuscript has been submitted at European Journal of Agronomy and is currently under revision.

Plant detection and counting from high-resolution RGB images acquired from UAVs: comparison between deep-learning and handcrafted methods with application to maize, sugar beet, and sunflower

Etienne David[1,*2], Gaëtan Daubige[2], François Joudelat[3], Philippe Burger[4],Alexis Comar[5], Benoit de Solan[2], Frédéric Baret[1]

[1]INRAe, UMR EMMAH, Avignon, France
[2]Arvalis – Institut du végétal, Avignon, France
[3]ITB – Institut Technique de la Betterave, Paris, France
[4]INRAe, UE Grandes Cultures Auzeville (GCA), Toulouse, France
[5]Hiphen, 22b Rue Charrue, Avignon, France

*** Correspondence:**
Corresponding Author
etienne.david@inrae.fr

Keywords: Plant counting, Deep-learning, Robustness, unmanned aerial vehicle, phenotyping. (Min.5-Max. 8)

1 Abstract

Measurement of the plant density is required for a number of applications since it drives part of the crop fate. The standard manual measurements in the field could be efficiently replaced by high-throughput techniques based on high-spatial resolution images taken from UAVs. This study compares several automated detection of individual plants in the images from which the plant density can be estimated. It is based on a large dataset of high resolution Red/Green/Blue (RGB) images acquired from Unmanned Aerial Vehicules (UAVs) during several years and experiments over maize, sugar beet and sunflower crops at early stages. A total of 16247 plants have been labelled interactively on the images. Performances of handcrafted method (HC) were compared to those of deep learning (DL). The HC method consists in segmenting the image into green and background pixels, identifying rows, then objects corresponding to plants thanks to knowledge of the sowing pattern as prior information. The DL method is based on the Faster Region with Convolutional Neural Network (Faster RCNN) model trained over 2/3 of the images selected to represent a good balance between plant development stage and sessions. One model is trained for each crop.

Results show that DL generally outperforms HC, particularly for maize and sunflower crops. A significant level of variability of plant detection performances is observed between the several experiments. This was explained by the variability of image acquisition conditions including illumination, plant development stage, background complexity and weed infestation. The image quality determines part of the performances for HC methods which makes the segmentation step more difficult. Performances of DL methods are limited mainly by the presence of weeds. A hybrid method (HY) was proposed to eliminate weeds between the rows using the rules developed for the HC method. HY improves slightly DL performances in the case of high weed infestation. When few images corresponding to the conditions of the testing dataset were complementing the training dataset for DL,

a drastic increase of performances for all the crops is observed, with relative RMSE below 5% for the estimation of the plant density.

2 Introduction

Plant density at emergence is a main yield component particularly for plants with reduced tillering or branching capacities such as maize, sugar beet and sunflower. The plant density at emergence is controlled by the seeding density and the emergence rate. Further, the seeding pattern defined by the distance between row and between plants influences the competition between plants and possibly with weeds. In addition to the estimation of plant density, the position of each plant can be documented to describe the local competitive environment [1]. For agronomical or phenotyping experiments, the plant density is mainly used to evaluate the quality of each microplot with consequences on the whole trial. It is also used by farmers to decide to stop spending resources to grow the crop in case of too low density or too much heterogeneity. While plant density is not directly governed by the genotype, but results from the seeding density, seed vigor and the emergence conditions, it is considered as an agronomical trait in some widely used ontology [2].

Plant density is currently mostly evaluated manually in breeding programs. Operators count plants in the field over a limited sampling area since this process is tedious, time-consuming, and therefore expensive. Consequently, this traditional method can lead to significant uncertainties due to the limited representativeness of the sampled area and possible human errors. Further, the position of plants is generally not documented because it would be even more tedious to measure each plant location.

#	Study	UAV	Crop	Object	Sessions	Localization	Method	Test independency
1	[3]	Yes	Sorghum	Head	1	Yes	ML	No[1]
2	[4]	yes	Wheat	Plant	5	yes	ML	No[1]
3	[5]	no	Wheat	Plant	several	yes	HC	Yes[2]
4	[6]	yes	Maize	Plant	1	yes	HC	yes[2]
5	[7]	yes	Sunflower	Plant	1	yes	HC	Yes[3]
6	[8]	yes	Agave	Plant	3	yes	HC	No
7	[9]	yes	Maize Sunflower wheat	Plant	6	no	HC (OBIA)	Yes[2]
8	[10]	yes	Thuja	Plant	3	yes	HC (OBIA)	Yes[2]
9	[11]	yes	Maize	Plant	2	yes	HC (OBIA)	No[1]
10	[12]	yes	Rapeseed	Plant	2	yes	HC (OBIA)	No[1]
11	[13]	Yes	Safflower	Plant	2	Yes	HC (OBIA)	No[4]
12	[14]	No	Wheat	Head	2	yes	DL	Yes
13	[15]	No	Maize	Plant	10	yes	DL	No[1]
14	[16]	Yes	Sorhgum	Plant	2	no	DL	No[1]
15	[17]	Yes	Wheat	Head	several	no	DL	Yes
16	[18]	Yes	Spinach	Plant	1	no	DL	No[1]
17	[19]	Yes	Maize	Head	2	yes	DL	No[1]
18	[20]	Yes	Sorghum	Head	2	yes	DL	No[1]
	This study	Yes	Maize Sugar beet Sunflower	Plant	27	yes	HC / DL	Yes

Table 1: Comparison of the different approaches used for plant and organ counting referenced in the literature. [1] random selection of samples for training and testing; [2]No proper calibration; [3]Calibrated with synthetic data; [4]Testing is made on two sessions, one session being already used for training

The recent technological advances of plant phenotyping solutions including Unmanned Aerial Vehicles (UAV), sensors, computers, and image processing algorithms, offer potentials to develop alternative methods to the manual counting. Several authors already reported accurate estimates of plant or organ counting and density from RGB images (Table 1). Plants or organ can be characterized either with

machine learning (ML) algorithms where standard local image features are extracted and a used in a supervised classification to identify the objects of interest (Guo *et al.*, 2018; Fernandez-Gallego *et al.*, 2019). Handcrafted (HC) methods rely on expert knowledge to compute the pertinent features and use them to identify the objects of interest. Most of them belong to the Object Based Image Analysis [9]–[13]. The identification process can be done based also on the expert knowledge [5]–[7] or by calibrating a statistical model over a training dataset [8]. More recently, approaches based on deep-learning (DL) have been proposed. The features are automatically extracted from the image and then used to identify and localize the individual objects of interest ([14], [15], [19], [20]). However, these features can also be used to estimate directly the density of objects through a regression [16]–[18]. Localization, is more popular (78% of the studies in Table 1) in plant phenotyping as it documents the sowing heterogeneity including missing plants, allowing to explore the competition between plants as outlined earlier. DL based methods are being common now to detect plant and organ and represent almost 30% of the localization studies (Table 1). Madec et al. [14] demonstrated that the Faster RCNN DL model [21] provides accurate localization of wheat ears with higher robustness than previous methods, including direct regression method. A higher heritability than that of manual counting was also reported. More recently, [19], [20] applied similar strategies to locate plant and organ from UAV images. DL applications to plant phenotyping are supervised learning methods, requiring large and diverse labelled datasets to converge to a generic solution. The recent progress in DL applied to detection/localization tasks beneficiated from the availability of large image collections such as ImageNet [22] and COCO Dataset [23] that are used to pre-train the DL model.

However, Geiros et al. [24] raised the overfitting risk and the resulting lack of robustness associated with most DL algorithms. They can reach excellent performances for datasets like those used for their calibration, while often failing when applied to cases different from the training dataset. In comparison, HC methods are based on expert knowledge which select the main features to identify the target objects. This reduces the risk of overfitting but can hardly account for all the specific cases. On the 11 methods listed (Table 1) that require a training dataset, only 3 [13], [14], [17] proposed a proper evaluation framework where the training and the test datasets do not come from the same acquisition sessions. This questions the accuracy, scalability and robustness of HC and DL methods that was investigated in the case of liver disease [25], but not for the plant detection problem within phenotyping applications.

The objective of this study is to compare a HC approach based on the knowledge of the sowing and plant patterns and a DL approach based on object detection to localize plants and count them. This study includes three species (maize, sugar beet and sunflower) observed with a RGB (Red Green Blue) camera aboard a UAV during 27 acquisition sessions with plants at different development stages few weeks after emergence. This study appears therefore to be the most comprehensive one on the subject (Table 1), while keeping always the training and test datasets as independent as possible. Further, we will also propose to combine the DL approach with expert knowledge from the HC one.

3 Materials and methods

3.1 Dataset

3.1.1 Experiments

The dataset used was acquired over maize, sugar beet and sunflower experiments from 2016 to 2019 in several experimental sites in France (Table 2). The sites cover a large diversity of agronomic modalities while managed with conventional tillage practices. However, some crop residues from the previous season can be observed on few microplots. Generally, few weeds were present in the microplots, except for some of them (Table 3). The sites include clay, brunisolic and limestone soil

types (Table 2) with a variety of surface roughness and moisture. The soil color varies from gray to brown due to soil type, surface aspect and illumination conditions. Each site included an ensemble of microplots corresponding to many genotypes from which 3 to 12 were selected to get approximately 600 plants (Table 3). Some sites were flown several times (Table 2), corresponding to several acquisition sessions. This allows to get a larger variation in the crop development stage during image acquisition (Table 3). For maize, a total of 51 microplots was available from 9 acquisition sessions (Table 3) with contrasted microplot size, row spacing (0.3-1.1m), and plant density (5.1-11.2 plt.m^{-2}). For sugar beet, a total number of 60 microplots was available from 9 acquisition sessions with microplot size, row spacing and plant density varying within a small range (Table 2). For sunflower, a total of 78 microplots was available from 9 acquisition sessions with a large variability of microplot size, row spacing, and plant density.

Crop	Site Name	Lat (°)	Long (°)	Year	Nb. sessions	Nb. microplots	Microplot width (m)	Microplot length (m)	Row spacing (m)	Plant density (plt.m^{-2})	Soil type
Maize	Menainville	47.9	1.4	2016	1	6	2.2	7.0	1.10	5.1	Clay
	Nerac	44.1	0.3	2016	1	8	1.6	7.0	0.80	8.5	Clay
	Villedieu	47.8	1.5	2016	1	6	0.9	11.0	0.30	19.9	Clay
	Thenay	47.3	1.2	2017	1	6	4.4	6.0	0.63	7.3	Clay / Flint
	Blois	47.7	1.2	2019	1	7	1.7	7.0	0.83	9.5	Brunisolic
	Castetis	43.4	-0.7	2019	1	5	2.8	4.0	0.70	11.2	Brunisolic
	Ermine	46.5	-1.0	2019	1	4	3.2	5.5	0.80	8.6	Limestone
	Selommes	47.7	1.2	2019	1	7	1.8	5.3	0.88	9.5	Brunisolic
	Pleinefougeres	48.5	-1.5	2020	1	2	3.2	11.0	0.80	7.7	Brunisolic
Sugar beet	Bucy	49.6	3.9	2017	2	7	1.4	6.2	0.45	11.1	Loam
	Charmont	48.3	4.1	2017	1	7	1.4	5.5	0.45	11.1	Limestone
	Etienne	49.2	4.3	2017	1	6	1.2	7.6	0.40	15.6	Limestone
	Memmie	48.9	4.3	2017	2	6	1.4	7.6	0.48	10.8	Limestone
	Charmont	48.3	4.1	2018	2	8*	1.4	5.5	0.45	11.4	Limestone
	Memmie	48.9	4.3	2018	1	6	1.4	7.6	0.45	11.4	Limestone
Sunflower	Rivière	43.5	1.5	2017	1	8	3.0	4.1	0.50	7.1	Clay
	Auzeville	43.5	1.5	2018	2	3	3.3	9.5	0.55	6.1	Clay
	Auzeville	43.5	1.5	2019	5	12	2.9	9.0	0.96	3.7	Clay
	Epoisses	47.2	5.1	2019	1	4	2.4	10.0	0.60	5.1	Limestone

Table 2. Characteristics of the crops for the several sites considered.

3.1.2 Acquisition and labelling details

Image acquisition was carried out by UAVs embarking three different RGB cameras including the Sony Alpha 5100, Sony Alpha 6000, both with a resolution of 6024x4024 pixel, and the Zenmuse X7 (DJI) in the case of Epoisses site in 2019 with a resolution of 6016 x 4008 pixels. The cameras were fixed on a two axes gimbal to maintain the nadir view direction during the flight. The camera was set to speed priority of 1/1250 s to limit motion blur. The aperture and ISO were automatically adjusted by the camera. The camera was triggered by an intervalometer set at 1Hz frequency corresponding to the maximum value allowed to record the RGB images in JPG format on the memory card of the camera. Flight altitude above ground varied between 20 to 50m to get a ground sampling distance (GSD) between 2 mm and 5 mm per pixel (Table 3). The flight trajectory was designed to ensure more than 70% overlap between images across and along tracks. Ground control points were placed in the

field and their coordinates were measured with a real-time kinetic GPS device ensuring an absolute centimetric accuracy of their position.

	Session_name	plant number	plot number	Stage	GSD (mm)	typical BB size (cm)	typical BB size (pixel)	Weed infestation	Blur
MAIZE	Selommes_2019_1	510	7	1	3.5	6.5	26	2	233
	Hermine_2019_1	542	4	1	3.5	7.8	31	1	79
	Thenay_2017_1	617	6	1	2.5	8.5	34	1	1149
	Castetis_2019_1	575	5	2	3.3	10.0	40	1	121
	Pleinefougeres_2019_1	504	2	2	3.5	11.5	46	0	39
	Blois_2019_1	579	7	2	3.3	12.3	49	1	346
	Menainville_2016_1	620	6	3	3.4	12.3	49	1	78
	Villedieu_2016_1	629	6	3	2.7	13.3	53	0	261
	Nerac_2016_1	594	8	3	4.0	15.0	60	0	37
	Total	**5170**	**51**						
SUGAR BEET	Memmie_2017_1	667	6	1	4.5	8.0	32	0	26
	Charmont_2018_1	556	7	1	4.2	11.5	46	0	93
	Memmie_2018_1	602	6	1	4.3	11.5	46	0	77
	Bucy_2017_1	634	7	2	5.3	12.8	51	0	25
	Memmie_2017_2	679	6	2	5.7	14.8	57	0	72
	Etienne_2017_1	635	6	2	4.5	16.0	64	0	27
	Charmont_2017_1	669	8	3	3.4	20.5	82	0	191
	Charmont_2018_2	647	8	3	4.1	20.5	82	0	102
	Bucy_2017_2	558	6	3	4.5	23.0	92	0	31
	Total	**5647**	**60**						
SUNFLOWER	Auzeville_2019_1	579	12	1	5.0	8.5	34	1	28
	Auzeville_2019_2	640	12	1	5.0	13.5	54	1	510
	Epoisses_2019_1	596	4	1	2.5	14.3	57	1	10
	Auzeville_2018_1	596	3	2	2.3	14.3	57	1	488
	Auzeville_2019_3	657	12	2	5.0	19.3	77	0	350
	Auzeville_2019_4	603	12	2	5.0	24.5	98	0	221
	Rivière_2017_1	634	8	3	5.2	25.0	100	2	42
	Auzeville_2018_2	560	3	3	2.6	27.5	110	1	1286
	Auzeville_2019_5	565	12	3	5	27.5	110	2	176
	Total	**5430**	**78**						

Table 3. Characteristics of each measurement sessions. For sugar beet, microplots from one session to another are the same. For sunflower the microplots considered change between sessions. The typical size of the BB for one session is computed as the square root of the mean area of all the BBs. The typical bounding box (BB) size in pixels is computed after up sampling the images at 2.5 mm resolution. The plant stage at the time of the session is quantified as: 1: early, 2: intermediate, 3: late. The correspondence with BBCH scale is provided as a table in the supplementary material section. The weed infestation is scored from 0 (no weed), from 0 (no weeds), 1 (less than 5% coverage), 2 (more than 5% coverage). The image blur is quantified by the average variance of the Laplacian: high blur results in low value of the variance of the Laplacian.

Agisoft Photoscan Professional software was used to align the images. The high overlap between the images and structure from motion algorithm permits to compute the position and orientation of the cameras. The pipeline described in Jin et al. [26] was then run to extract from each image the portion

corresponding to the contained microplots. Using the original images avoids the possible distortions and artefacts observed in the orthomosaïque. Several extracts may represent the same microplot viewed from different positions of the UAV [27]. For each microplot, the sharpest extract that contained the whole microplot is selected. For each session, a few microplots were selected for labelling (Table 2). Approximately 600 plants per session were labelled to ensure consistency across sessions which resulted in a total of 16247 labelled plants. Images were rescaled to match the best available GSD (2.5 mm, Table 3). This was necessary to control the apparent size of object, which can make the Deep Learning methods fail. Then all images were labelled using the coco-annotator tool [28] The labelling consisted in drawing a bounding box (BB) around each plant. Six different operators contributed to the labelling. The labelling from one operator was always reviewed at least once by a different operator. The typical size of the BB for one session (Table 3) was computed as the square root of the mean area of all the BBs.

The plant development stage during the acquisition sessions was scored into three relative levels, where stages 1 ,2 and 3 correspond respectively to early (few days after emergence), intermediate, and late stages (leaves start to fill the gap between plants). The correspondence between the stages for each crop, and their BBCH scale is presented in Table S1. The level of weed infestation (Table 3) was also visually evaluated from 0 (no weeds), 1 (sparse presence of weeds), 2 (infestation). The level of blurriness for each session (Table 3) was evaluated by calculating the average variance of the Laplacian [29].

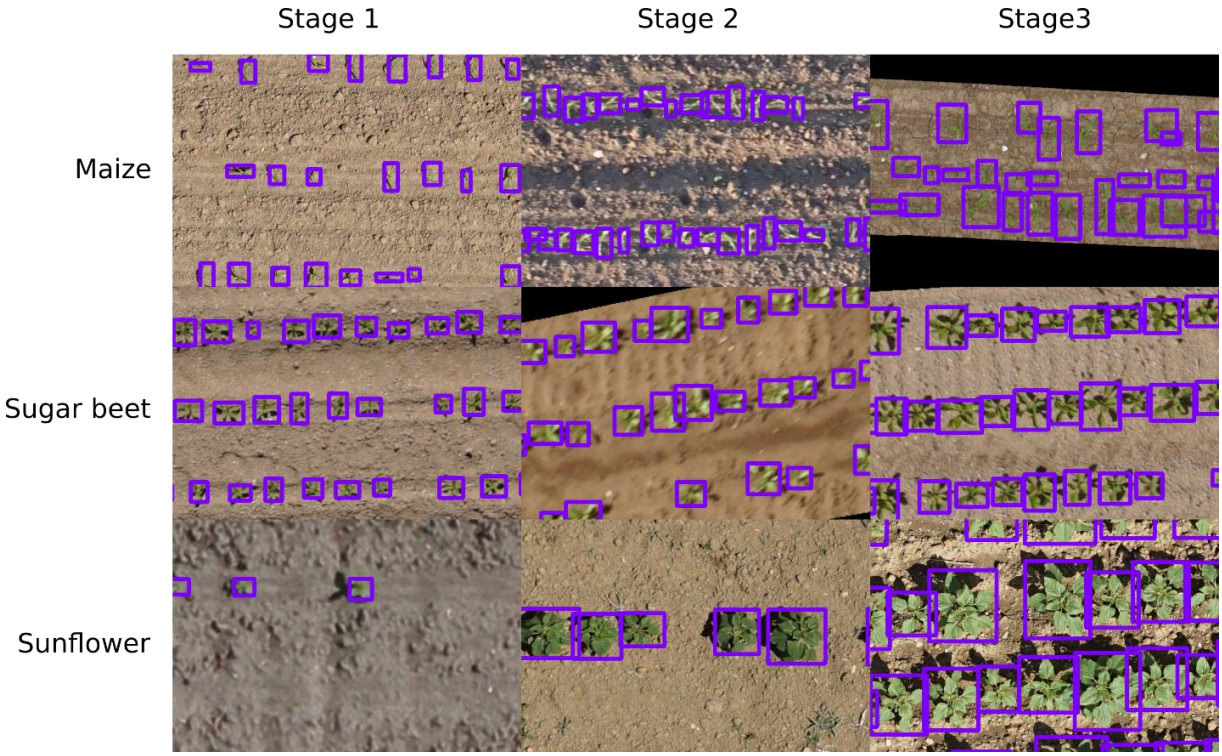


Figure 1: Samples of images for the three-development stage. All images were resampled to 0.25mm.px^{-1}. The bounding boxes were drawn interactively around the plants.

3.2 Plant detection methods

3.2.1 Handcrafted method

The method developed here is based on several assumptions: (1) the plants are green and can be accurately separated from the background; (2) plants are sown in rows relatively evenly spaced and parallel; (3) the weeds are mainly located in between the rows and are not too dominant; (4) plants are relatively evenly spaced on the row and are not too variable in shape and size. The method first extracts each single row and then identifies each individual plant on the row. All the parameters of our HC method are expressed in relative value to the row or plant spacing, to allow adaptation to a larger number of sowing patterns. This makes our method scalable to all our experimental conditions across the three species (Table 2 and table 3). The values of the parameters were set based on reasonable assumptions and were not calibrated on a dataset.

3.2.1.1 Row extraction

The original RGB images are first transformed into a black and white one (BW) using the excess green index (ExG =(2G-B-R)/G) where R, G, B correspond respectively to the red, green and blue colours of the original image [30]. Pixels are then assigned to the green (1) or background (0) classes using the ExG threshold value defined with the Otsu algorithm for each session [31].

The Hough transform [32] is used to identify the main alignments corresponding to the rows and find their orientation. The image is then rotated to display the rows horizontally (Figure 2). The number of green pixels in each line is computed to obtain a profile of green pixels across the rows. The peaks of the green pixel profiles are localized using the prior knowledge on row spacing (*Row_spacing_prior*) to prevent finding unexpected peaks between rows. The prior knowledge of the number of rows per microplot (*Row_number_prior*) is also used when identifying the peaks. The prior values of row and plant spacing are not always known precisely. Therefore, the row extraction pipeline (Figure 2) provides also updated and more accurate values of *Row_spacing_prior* for each session. Finally, each row is extracted using the fine-tuned value of the row width.

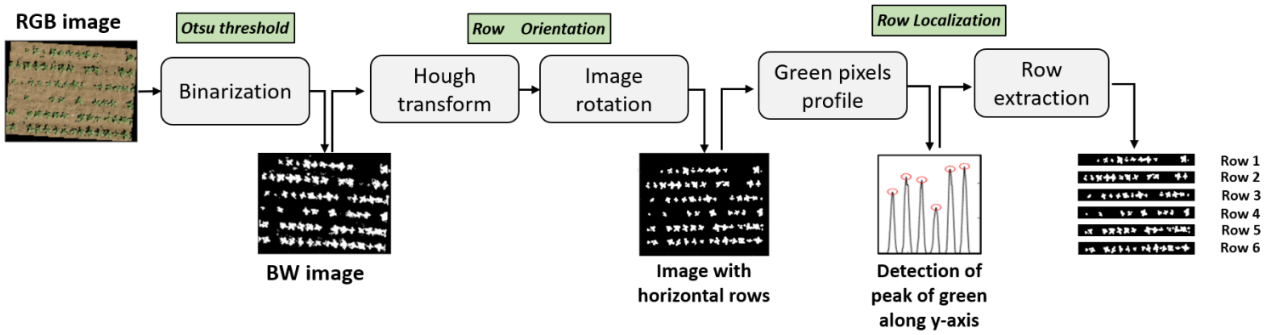


Figure 2. Flowchart of the rows extraction process from the original RGB image.

3.2.1.2 Plant identification with an object-based method

After the row extraction, the algorithm individualizes the objects (groups of connected pixels) in the image and classifies them as plants or weeds. Weeds are eliminated based on the distance to the row center. If the centroid of an object is located at a distance larger than a threshold value (*Minimum_distance_to_row)*, it is considered as a weed. The threshold value is expressed in relative value to the row spacing and set to 0.25 (Table S2). Objects with dimensions along the row direction larger than the *Plant_spacing_prior* value (Table 2) are expected to include several plants. The number of plants contained in these big objects is derived from the number of peaks observed when summing

the green pixels along the row direction, where a peak may correspond to a plant position. Further, the number of plants found by the number of peaks is crosschecked with the expected number of plants computed by dividing the extension of the object by the *Plant_spacing_prior* value. Results are illustrated in Figure 3 for the two objects on the right of the bottom row.

Finally, some objects may be located too close together to be considered as separate plants because these objects correspond to several parts of the same plant. Figure 3 illustrates it with the second plant starting from the left on the top row, where a leaf and the main plant are separated. If the distance between the centroids of the closest object is smaller than the maximum acceptable distance, *Big_plants_tolerance x Plant_spacing_prior*, the two objects are merged as a single plant. Table S2 in the supplementary materials presents the value used for each parameter. The centroid (center of mass of the object), and the bounding box (smallest rectangle that contains all object's pixels) of the objects are finally computed.

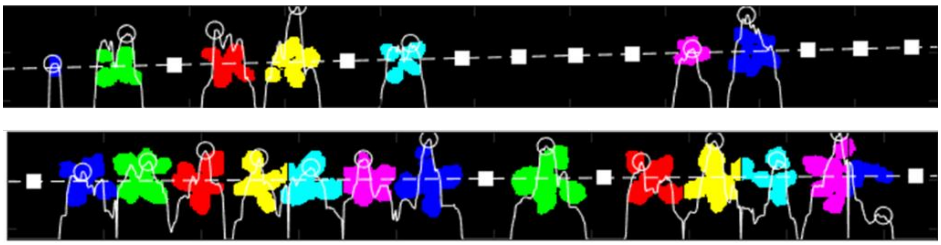


Figure 3. Typical output of the HC algorithm illustrated for two sugar beet rows. The dashed white line indicates the row. The white curve represents the profile of number of green pixels perpendicular to the row, with peaks identified by a circle. The object-based method is illustrated by the colors assigned to each identified plant. Note that big objects have been split into individual plants (bottom row, the four last plants) and isolated plant parts have been reconnected to form a single plant (top row, fourth plant starting from the left). The white squares correspond to the position of missing plants.

3.2.2 Deep-learning method

3.2.2.1 Model architecture

An object detection method was selected to predict the bounding box around each plant. This information can then be used to derive more traits to characterize every individual plant. Object detection is a fast-growing area within DL techniques since the emergence of networks such as R-CNN (Regions with Convolutional Neural Network , [33]) or SSD (Single Shot Detector, [34]). Most DL object detection models fall into one-stage or two-stage models. In the one-stage model, the object is localized and categorized in a single step. In the two-stage model, a first stage detects possible objects, and a second stage categorizes them. The Faster-RCNN two-stage model [21] is used because it performs well in the context of plant phenotyping. Madec et al. [14] used it successfully for counting wheat heads. It allows also to analyze the nature of the possible errors by visualizing them.

Faster-RCNN can be implemented in many forms which can influence the final results. We use the implementation made by the mmdetection library [35] .It contains many detectors, and is written upon PyTorch [36]. The default implementation of the library is used and contains a Feature Pyramidal Network (FPN) [37], which differs from the original paper [21]. It is used to provide object proposition at different scales. A ResNet-34 model [38] was used as the backbone network because it offers a good compromise between accuracy and speed of training. The backbone extracts the deep features which are used by the Region Proposal Network (RPN) to detect potential objects which are then classified

as crop or background. All other architectural details are given in the code (https://github.com/EtienneDavid/plants-counting-detection) . We also choose to train one model by crop as preliminary tests show lower performances when mixing the three crops.

3.2.2.2 Pre-processing and data augmentation

The input image size of the network is set to 512 x 512 pixels to match memory constraints during training. However, images from the microplots are larger. A preprocessing step first splits them randomly into patches of 512 x 512 pixels. For each session in the training dataset, 100 patches were randomly selected which results in a total of 900 patches to train the model for each crop over the nine available sessions. Randomly sampled patches provide more diversity than evenly sampled ones. During the training process, data augmentation is applied to extend the diversity of images. The complete data augmentation pipeline is a set of geometric distortions (Random rotation, Random Translation, Random Shear), blur (Gaussian Blur), noise (Gaussian noise) and colorimetric augmentation (Random hue value, Random contrast). All data augmentation details are given in the code. Once trained, the model is applied to all the patches. Predictions from the overlapping patches are finally merged together by using the Non-Max-Suppression algorithm [39] with an Intersection over Union (IoU) threshold of 0.70.

3.2.3 Hybrid method

DL methods detect individual plants based on many features automatically extracted while HC methods exploit expert prior knowledge on the sowing pattern to eliminate plants located at a non-expected position between rows. We propose therefore a hybrid method that combines the benefits of both HC and DL ones. The DL method is first applied to detect plants. Then, the HC method presented earlier is used to identify the row position and eliminate all remaining weeds corresponding to plants with centroids located at a larger distance to the row than a threshold value *distance_to_row* (Table S2).

3.3 Evaluation strategy for plant detection

3.3.1 Strategies for training and evaluation

Detection models were developed and evaluated independently for each crop. DL method requires an extensive training dataset that should represent the expected diversity of situations. Due to the limited number of labelled images, two strategies are defined: “Out-Domain” and “In-Domain”. “Out-Domain” is the more rigorous strategy where the performances of the DL method are evaluated over sessions not used during the training process. For each crop and each stage, two sessions were used for training and the remaining one for testing. This allows to balance the stages between the training and testing datasets. A three-fold cross-validation strategy that exploits all sessions while providing relatively independent test cases is used. Three different models were trained for each crop using six sessions, representing about 3800 plants, and tested on the remaining three sessions representing around 1900 plants. The “In-Domain” strategy is based on adding few images randomly selected in the testing datasets to the training dataset. It aims at reducing possible lack of representativeness in the training dataset. The same three-fold cross-validation process was used for each crop, except that 1/3 of the 600 plants used previously as testing datasets were added to the training dataset. The remaining 2/3 images (400 plants) are used to evaluate the performances of the models for each crop. The same test dataset (1200 plants corresponding to the 400 test plants for each of the three test sessions) is finally used to compare the Out-domain and In-domain approaches.

3.3.2 Evaluation metrics

Detection

The "Centroid matching strategy" (C_MS) is used to evaluate whether a plant was correctly detected. The C_MS is based on the distance between the centroids of the plants. If the distance between centroids of a detected plant and the closest labelled one is smaller than *Plant_distance_prior* / 2 it is considered as true positive (TP). Otherwise, it is a false positive (FP). If a labelled plant has no detected plant within a distance smaller than *Plant_distance_prior* / 2, it is a false negative (FN).

The plant detection performance was quantified per session with the terms of the confusion matrix normalized by the number of labelled plants (TP+FN) for easier comparison between crops and stages, which correspond to rates of TP (TPR), FP (FPR) and FN (FNR). The accuracy is also used, defined as TP/(TP+FN+FP). DL method produces a confidence score for each predicted BB. A box is considered as a prediction for the DL and HY methods if its score is above 0.5.

Plant density

Plant density (PD) was calculated by dividing the number of plants in the microplot by its area. The area is computed as the number of rows multiplied by the row spacing and the row length. The relative root mean square error (rRMSE) is used to compare the estimated and the reference PD values and assess the accuracy of the method. The accuracy levels were split into four classes to better assess the robustness of the method. A rRMSE<5% was considered as good, between 5%<rRMSE<10% as satisfactory, between 10%<rRMSE< 20% as poor, and rRMSE>20% as very poor. The percentile of microplots belonging to each class was therefore used to evaluate the robustness of the methods.

4 Results and discussion

4.1 DL and HY methods detect better plants than the HC one

Detection performances are very different depending on the crops (Table 4 and Figure 4). Detection of maize plants appears difficult for the three methods and particularly for HC with a low TPR and a high FNR (Table 4). This is mostly explained by the small size of the plants that overlap, resulting into groups of overlapping plants that are interpreted as a single plant (Figure 5b), or to poor threshold values determined by the Otsu method for the green segmentation used in the first step to identity objects (Figure 5a). However, a high FNR is also observed for the first development stage with the HC method, due to the poor quality of the green segmentation where background artifacts such as small rocks or crop residues were interpreted as plants (Figure 5g). A large variability between the three instances of the three-fold cross validation is observed for this early stage (Figure 4) due to the variability in image quality. Marginal differences are observed between DL and HY methods. They both show relatively balanced FPR and FNR. FPR is mostly explained by possible confusion between plants and their shadows or soil artifacts (Figure 5c) while FNR is explained by the small size of the plants that are difficult to detect (Figure 5d). This results into accuracy values between 0.77 to 0.80 with little variation between stages (Table 4). However, a larger variability across the three instances of the three-fold cross validation is observed for the late stage (Figure 4).

Crop	Stages	N	TPR			FPR			FNR			Accuracy		
			HC	DL	HY	HC	DL	HY	HC	DL	HY	HC	DL	HY
Maize	1	1669	0.61	0.88	0.86	0.27	0.12	0.07	0.39	0.12	0.14	0.56	0.79	0.80
	2	1658	0.70	0.92	0.92	0.03	0.18	0.16	0.30	0.08	0.08	0.68	0.78	0.79

	3	1930	0.70	0.88	0.86	0.05	0.15	0.14	0.30	0.12	0.14	0.68	0.77	0.77
Sugar beet	1	1825	0.95	0.98	0.98	0.04	0.01	0.01	0.05	0.02	0.02	0.93	0.97	0.97
	2	1948	0.95	0.99	0.99	0.01	0.03	0.03	0.05	0.01	0.01	0.93	0.97	0.97
	3	1874	0.94	0.99	0.99	0.06	0.04	0.04	0.06	0.01	0.01	0.88	0.95	0.95
Sunflower	1	1603	0.80	0.87	0.86	0.17	0.06	0.04	0.20	0.13	0.14	0.75	0.80	0.81
	2	1856	0.82	0.94	0.94	0.15	0.08	0.07	0.18	0.06	0.06	0.72	0.87	0.88
	3	1759	0.86	0.97	0.97	0.42	0.43	0.21	0.14	0.03	0.03	0.61	0.74	0.81

Table 4: Terms of the confusion matrix for the three methods the three crops, and the three stages. True Positive Rate (TPR), False Positive Rate (FPR), and False Negative Rate (FNR) are displayed. N is the true number of plants (N=TP+FN). Green color corresponds to good metrics values (high for TPR, low for FPR and FNR), and red for poor metrics values (low for TPR, high for FPR and FNR).

Detection of sugar beet plants appears to be much easier, with performances similar between the three methods. The sugar beet crops better verify the assumptions described in 3.2.1. The plots were not infested by weeds (Table 4), which seems to be an important explanation for the success of all methods. A small FPR is observed for the three methods, particularly for the latest stage, which explains the decrease in accuracy (Table 4). This is due to difficulties when plants are overlapping (Figure 5e). Slightly higher FNR is observed for HC corresponding to non-detected plants in the case of small plants and image of poor quality. This is also observed with DL for the very early stages (Figure 5f). The variability across the three instances of the three-fold cross validation is also small (Figure 4). Marginal differences are observed between DL and HY methods mostly because of the good control of weeds.

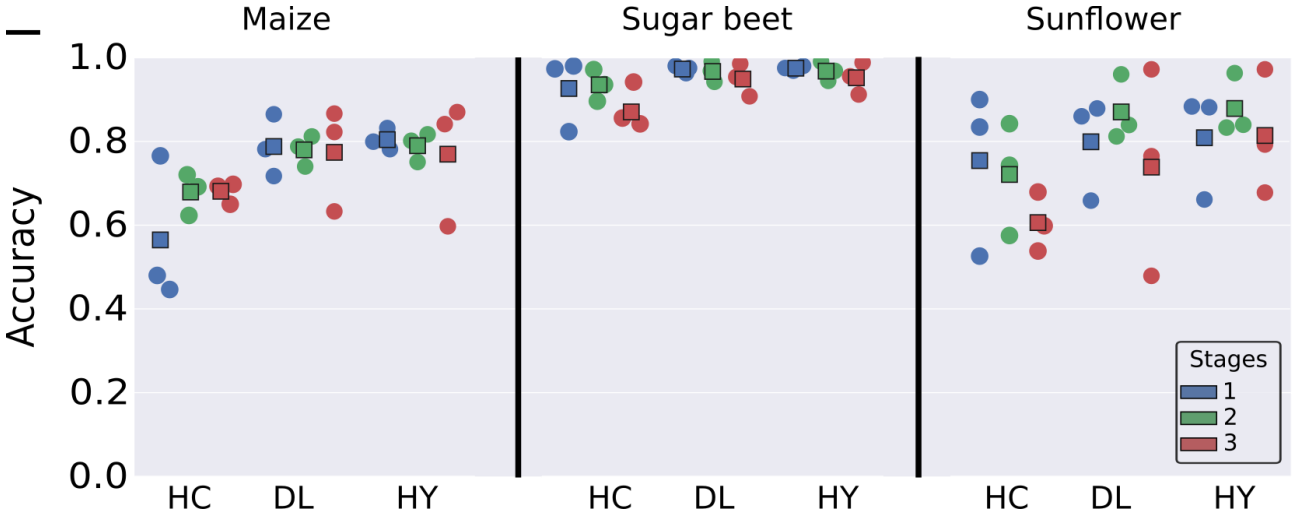


Figure 4: Accuracy for all methods and crops. For each crop and method, the stages are represented by a specific color. Each point corresponds to a test session used in the three-fold validation process. The squares represent the average of the three points.

Detection of sunflower plants shows accuracy values intermediate between maize and sugar beet (Table 4 and Figure 4). The HC shows lower TPR and higher FPR and FNR as compared to DL and HY. In the late stage, the HC shows very high FPR corresponding to problems of plant separation when they are overlapping. Further, the weeds close to the row line are not well eliminated and confounded with plants (Figure 5g). Similar problems are observed for the DL method, with weeds confounded

with the crop. However, the HY methods allows to eliminate part of the weeds that are located in between rows (Figure 5h). and HY shows high and similar TPR (Table 4). However, a high FPR is also observed for the first stage with the HC method, due to the poor quality of the green segmentation where background artifacts, such as small rocks or crop residues, were interpreted as plants (Figure 4g). Conversely, high FPR are observed for the late stage where DL shows difficulty to detect plants in a group of overlapping ones and confounds weeds with the crop. A large variability between the three instances of the three-fold cross validation is observed for sunflower (Figure 4). It is explained by a high degree of heterogenety in the microplots and between them, as well as between sessions.

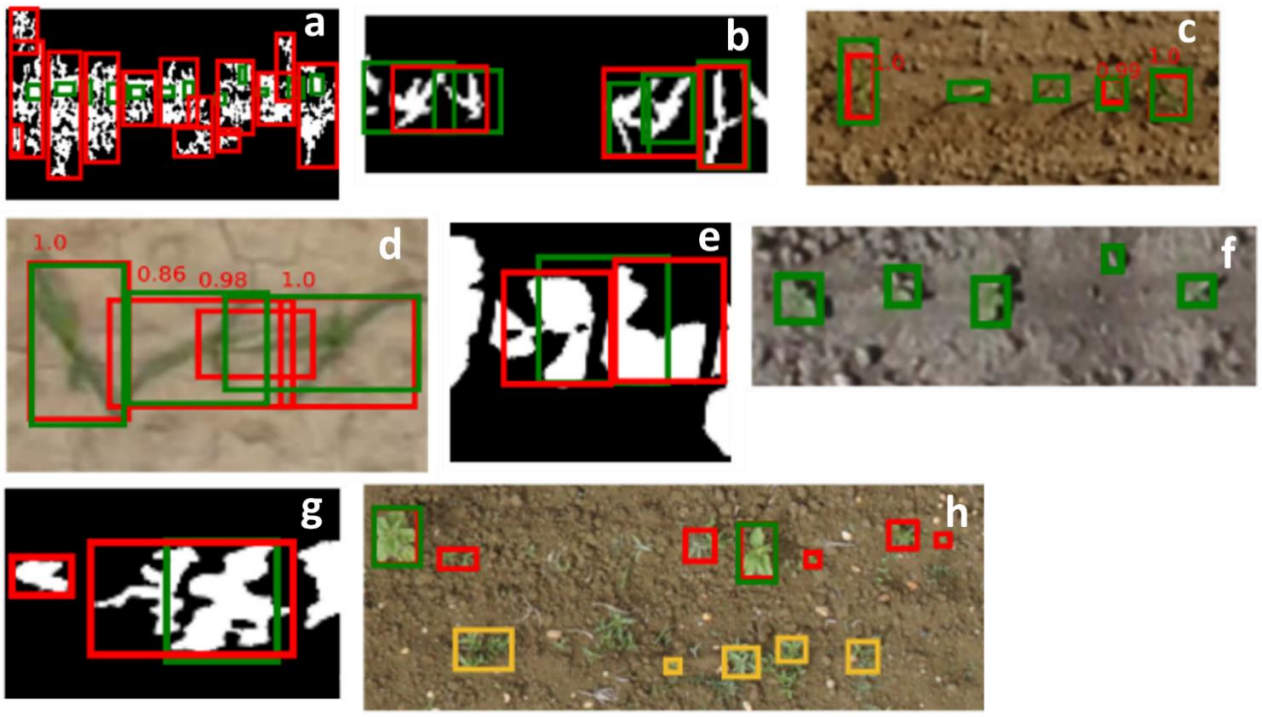


Figure 5: Possible detection errors for HC and DL methods. The green BBs correspond to the labelled plants. The red BBs correspond to the detected plants and yellow boxes correspond to weeds detected as crop. RGB images are displayed for the DL method. BW images are displayed for the HC method. a, b, c, d corresponds to maize, e, f, to sugar beet and g, h to sunflower.

Image quality appears therefore mandatory for HC methods to get a good segmentation. The HC methods appears also limited to eliminate weeds on the rows and to separate efficiently the overlapping plants. DL methods are similarly limited in separating crops from weeds, with confusions made mostly on unseen type of weeds (Figure 5h). However, the HY methods allows to eliminate part of the weeds. The DL methods also show some difficulties in detecting plants when they are small or when their shadows or other soil artifacts such as cracks are present. Nevertheless, the DL methods outperform the HC ones in most cases.

Tests were further conducted to evaluate the impact of the four qualitative factors (crop type, development stages, weeds, and soil type) using the p-value computed from a variance analysis. Results show (Table 5) that crop-type is an important factor (p_value smaller than 0.05) for HC and HY, while weeds are important for HC and DL, and soil-type for HC. However, the low number of examples (27 sessions in total), and the non-evenly distribution of the several factors (for instance most examples of

high levels of weed infestation are found in sunflower sessions only) prevents from drawing final conclusions. The impact of the four quantitative factors (sowing density, plant size, original resolution, and blurriness) were also evaluated using a Pearson test. It reveals (Table 5) that no factors appear significant (p-value smaller than 0.05), while the lowest p-values are observed for the sowing density and plant size that are closely related to the crop type.

Table 5: p-values computed from an ANOVA for the qualitative factors and Pearson test for the quantitative factors.

Factors	Type	HC	DL	HY
Crop type	qualitative	0.009130**	0.127550	0.032050**
Development stage	qualitative	0.857810	0.479530	0.643620
Weed infestation	qualitative	0.032610**	0.001600**	0.074540
Soil type	qualitative	0.026430**	0.781090	0.830650
Sowing density	quantitative	0.067379	0.076542	0.091679
Original resolution	quantitative	0.905626	0.572383	0.616534
Plant size	quantitative	0.791437	0.064765	0.211019
Blurriness	quantitative	0.111743	0.562775	0.501980

4.2 Plant density is better estimated with DL and HY methods

The HC method provides the poorest performances for maize plant density estimation, with rRMSE generally higher than 0.2 (Figure 6), which is consistent with the poorer detection performances (Figure 4). Image acquisition during the early stages tends to degrade the performances conversely to what was observed for the detection (Figure 5). This may be explained by the unbalance between false positives and negatives observed for the early stages (Table 4). Marginal differences are observed between DL and HY methods for maize where weeds were not the main issue.

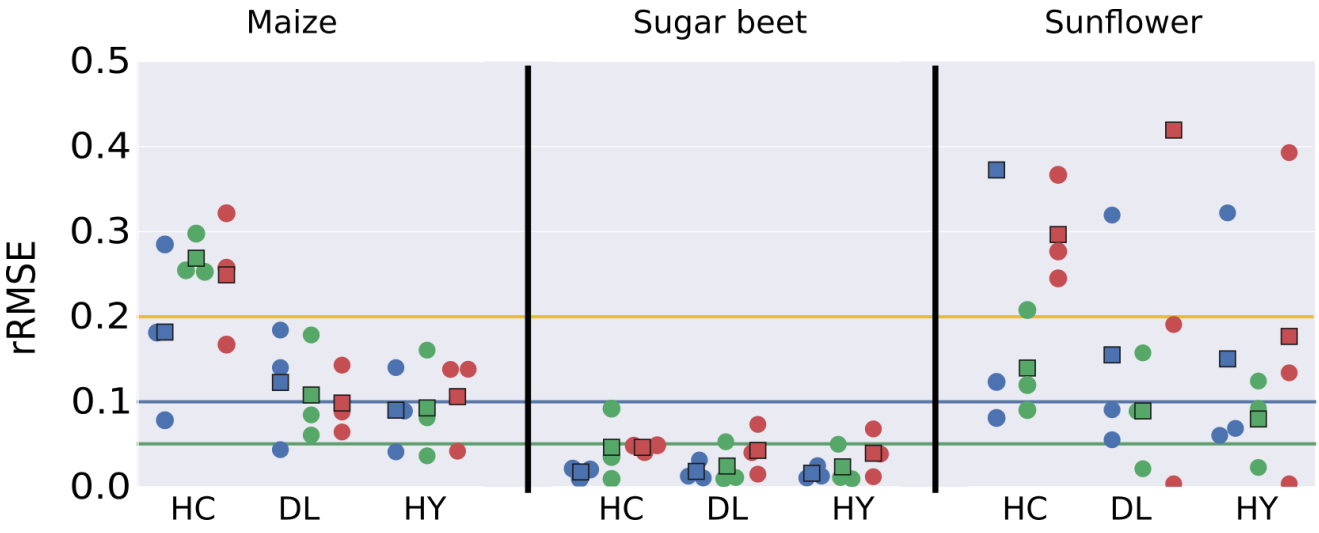


Figure 6: rRMSE for plant density estimation for all methods crops, and stages. Results obtained over the testing dataset. For each crop, method and stage, the three instances (corresponding to three testing sessions) of the three-fold cross validation process are displayed as colored dio sks,

while the corresponding average is represented by a colored square. Colors correspond to stages. The rRMSE threshold values to

All the methods reach good performances (rRMSE<0.05) for sugar beet, with even better performances for the two first stages when plants are easily identified and weeds not too developed (Figure 6). The poorer detection performances noticed earlier for HC (Figure 4) do not impact the density estimation because the FPR is well compensated by the FNR.

Sunflower shows more variability between sessions and stages, with rRMSE around 0.1 for the intermediate development stage showing better performances than the early one and moreover than the late one (Figure 6). The models for sunflower are very poor for the session 3_auzeville_2019_5 (Figure 6), mainly because of weed infestation. DL performs better than HC while HY improves marginally the performances for the two early stages, but significantly for the late stage where significant weed infestation was observed.

Overall, our results show lower performances than those of the studies where the training and testing datasets were not independent. For maize detection accuracy between 0.93 and 0.96, and relative counting error around 1.5%, were reported [11], [15] while none of our methods achieve such performances. Similar range of results are obtained on rapeseed (counting error of 6.83%) [12], or safflower with rRMSE approximately under 5% [13]. However, our results with DL and HY are comparable to studies keeping the training and test datasets independent; on maize Gnädinger and Schmidthalter [6] reports a counting error of +/- 15%. The HC approach applied when its main assumptions are verified performs well and comparably to DL.

4.3 Adding few images from the test domain improves drastically the DL performances

The performances of DL methods are closely related to the number of images used in the training dataset and their representativity of the possible situations [24]. DL method works very well for sugarbeet where all the images were relatively similar across sessions for each development stage. However, the acquisition conditions were quite different from the ones experienced in the other sessions for the sunflower on Epoisses_2019_1, explaining why the DL models had more difficulties to detect plants for this session. The “Out-domain” strategy used previously was compared here to the “In-domain” one where 1/3 of the images of the initial testing sessions were used to finetune the model. Performances are evaluated on the remaining 2/3 images of the initial testing sessions to keep some independence between the training and test datasets.

Note first that the plant density estimation performances (Figure 7) evaluated on a limited test data set (1200 images) are very consistent with the ones presented previously over the full test dataset including 1800 images (Figure 6). Results show that the additional images used in the training process and having similar characteristics as those in the testing dataset decreased significantly the rRMSE for all crops (Figure 7). This outperforms the marginal gain observed with the HY method on few sessions. Training with the In-domain strategy reduces the variability of performances across sessions. The 5% rRMSE value is reached for all crops except maize, where performances are anyway close to this target. Plant overlapping and the small leaf size makes the DL method for maize more challenging. However, there are still some outliers for Maize and Sunflower, corresponding to Pleinefougeres_2019_1 and Epoisses_2019_1 sessions. The images of these two sessions are highly blurred (Table 3) explaining most of their poor detection performances. A large part of this performance can be attributed to the elimination of almost all weeds by the DL methods, without the need of the HY correction, which have learned the pattern of the weeds, instead of relying on the location, and a better recognition of the plants.

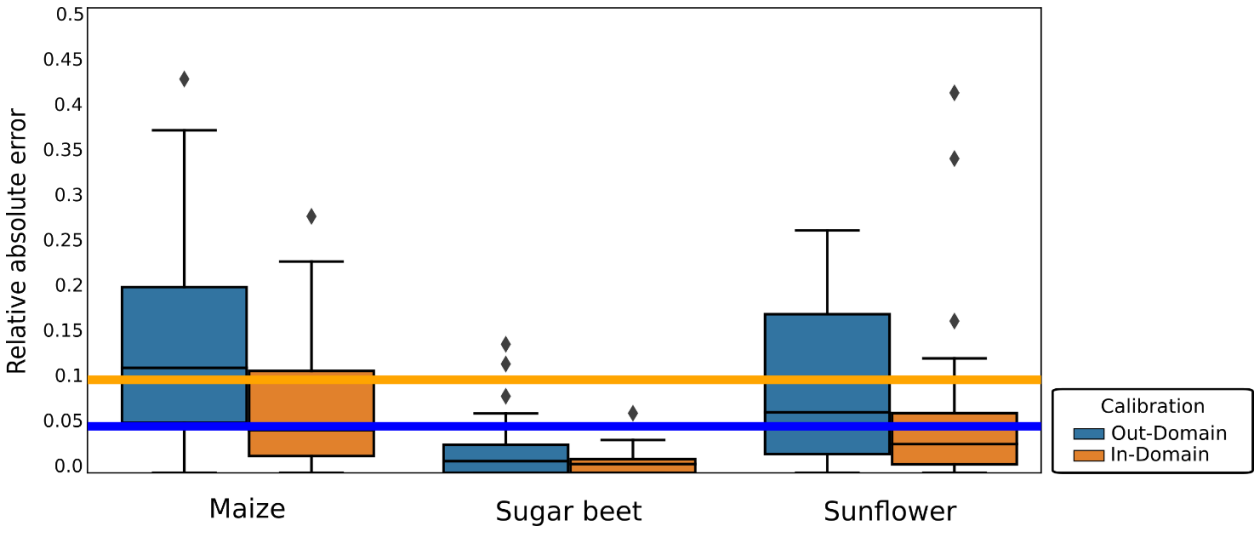


Figure 7: Distribution of relative absolute error for each microplots for the Out-Domain and In-Domain approaches for DL. Box-plot representation where the black horizontal bar represents the median, the box represents ±25%, the whiskers while the whiskers extend to the the lowest (highest) data point still within 1.5 interquantile range of the lower (upper) quartile. Diamonds are outliers. 1 outlier for Out-domain Maize and 3 outliers for Out-Domain Sunflower are above 0.5 and are not presented on the graph.

Our results demonstrate that active learning techniques [39] could greatly improve DL model performances for these new sessions. A small sample of images coming from the new sessions to be processed have to be labelled to complement the training dataset, but more than quantity, it is uniquely due to the diversity: only 40m² of maize or sugarbeet, and between 50 and 100m² of sunflower have been added to the training dataset, leading to a dramatic increase of the performances which cannot be attributed only to the dataset size increase. These results demonstrate the importance of having a proper design of DL training dataset when proposing a new trait to get robust estimates as required by agronomists, breeders, and farmers.

Our results are consistent with those of previous studies: detection and density estimation performances are generally lower when the training and the test datasets are independent, i.e not coming from the same measurement sessions. Fernandez-Gallo [ref] report a rRMSE below 5%, Madec et al. [ref] report a rRMSE of 15% on an independent test set. Similar drop in performances seems to happen in maize when comparing the results of Varela et al. (counting error of 1.5%) to those of Gnädinger and Schmidhalter (counting error of +/- 15%). The generalization potential of DL methods is high, requiring including more diverse situations in the training dataset at the expense of the tedious and expensive interactive labelling process. However, alternative techniques could be used to bypass this limitation, including data sharing between several organizations as this was done for the head counting problem (David et al., 2020). Data augmentation [40] could also improve greatly the generalization performances of DL methods. It would consist in manipulating the quality of the images, while creating synthetic images where a wide diversity of plants and weeds would be placed over different backgrounds with variation in the development stages and sowing pattern.

5 Conclusion

This study was based on a comprehensive dataset covering three main crops, several growth stages and acquisition conditions. It will be open to the community on Zenodo

(https://zenodo.org/record/4890370) to be possibly used as a benchmark for plant counting and detection from RGB images acquired from UAVs. Our results show that when the main assumptions on the sowing patterns are verified, simple HC methods can reach good enough performances to be used for applications as it was observed here for sugar beet. However, state-of-the art Deep Learning methods generally outperform the HC ones. Nevertheless, due to the large heterogeneity in terms of background, plant shape and phenological stages encountered across the wide collection of images considered, we demonstrated that the performances of the DL methods largely depend on the training and test datasets used. When the training domains used for the DL method are fully independent from the testing ones, the overall performances are reduced due to the failure of the model in a number of test cases poorly represented in the training dataset. Conversely, when adding few examples of images representative of the test domain, the performances increase drastically to reach those reported in most studies where training and test domains are not differentiated. Important gain in robustness could therefore be reached by including in the training dataset few images coming from the inference domains. Alternatively, a better understanding of the factors of variability between domains could constitute the basis to generate efficient data augmentation techniques that may even include synthetic images. An extended version of the dataset is needed to conclude on the main factors of error on plant counting with UAV. The hybrid method proposed to better eliminate weeds could be replaced efficiently by including images of the canopy where weeds were artificially incrusted.

6 Acknowledgments

The work 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. Many thanks to the dataset contributors:

- Arvalis (maize) : Menainville, Nerac, Villedieu, Thenay, Castetis
- Hiphen Plant (maize) : Blois, Selommes, Ermine, Pleinefougère
- Institut Technique de la Betterave (ITB) : All sugarbeet datasets
- Terres Inovia (sunflower) : Epoisses
- INRAe (sunflower): Auzeville, Rivière

7 References

[1] R. J. Godwin et P. C. H. Miller, « A Review of the Technologies for Mapping Within-field Variability », *Biosystems Engineering*, vol. 84, nº 4, p. 393-407, avr. 2003, doi: 10.1016/S1537-5110(02)00283-0.

[2] R. Shrestha *et al.*, « Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice », *Frontiers in Physiology*, vol. 3, p. 326, 2012, doi: 10.3389/fphys.2012.00326.

[3] W. Guo *et al.*, « Aerial Imagery Analysis – Quantifying Appearance and Number of Sorghum Heads for Applications in Breeding and Agronomy », *Frontiers in Plant Science*, vol. 9, p. 1544, 2018, doi: 10.3389/fpls.2018.01544.

[4] J. A. Fernandez-Gallego *et al.*, « Automatic wheat ear counting using machine learning based on RGB UAV imagery », *The Plant Journal*, vol. 103, nº 4, p. 1603-1613, août 2020, doi: 10.1111/tpj.14799.

[5] T. Liu, W. Wu, W. Chen, C. Sun, X. Zhu, et W. Guo, « Automated image-processing for counting seedlings in a wheat field », *Precision Agriculture*, vol. 17, nº 4, p. 392-406, 2016, doi: 10.1007/s11119-015-9425-6.

[6] F. Gnädinger et U. Schmidhalter, « Digital Counts of Maize Plants by Unmanned Aerial Vehicles (UAVs) », *Remote Sensing*, vol. 9, nº 6, Art. nº 6, juin 2017, doi: 10.3390/rs9060544.

[7] E. Jacopin *et al.*, « Using Agents and Unsupervised Learning for Counting Objects in Images with Spatial Organization », in *Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, 2021, p. 688-697. doi: 10.5220/0010228706880697.

[8] G. Calvario, T. E. Alarcón, O. Dalmau, B. Sierra, et C. Hernandez, « An Agave Counting Methodology Based on Mathematical Morphology and Images Acquired through Unmanned Aerial Vehicles », *Sensors*, vol. 20, nº 21, 2020, doi: 10.3390/s20216247.

[9] J. Torres-Sánchez, F. López-Granados, et J. M. Peña, « An automatic object-based method for optimal thresholding in UAV images: Application for vegetation detection in herbaceous crops », *Computers and Electronics in Agriculture*, vol. 114, p. 43-52, juin 2015, doi: 10.1016/j.compag.2015.03.019.

[10] Josue Nahun Leiva, James Robbins, Dharmendra Saraswat, Ying She, et Reza J. Ehsani, « Evaluating remotely sensed plant count accuracy with differing unmanned aircraft system altitudes, physical canopy separations, and ground covers », *Journal of Applied Remote Sensing*, vol. 11, nº 3, p. 1-15, juill. 2017, doi: 10.1117/1.JRS.11.036003.

[11] S. Varela *et al.*, « Early-Season Stand Count Determination in Corn via Integration of Imagery from Unmanned Aerial Systems (UAS) and Supervised Learning Techniques », *Remote Sensing*, vol. 10, nº 2, 2018, doi: 10.3390/rs10020343.

[12] B. Zhao *et al.*, « Rapeseed Seedling Stand Counting and Seeding Performance Evaluation at Two Early Growth Stages Based on Unmanned Aerial Vehicle Imagery », *Frontiers in Plant Science*, vol. 9, p. 1362, 2018, doi: 10.3389/fpls.2018.01362.

[13] J. C. O. Koh, M. Hayden, H. Daetwyler, et S. Kant, « Estimation of crop plant density at early mixed growth stages using UAV imagery », *Plant Methods*, vol. 15, nº 1, p. 64, juin 2019, doi: 10.1186/s13007-019-0449-1.

[14] S. Madec *et al.*, « Ear density estimation from high resolution RGB imagery using deep learning technique », *Agricultural and Forest Meteorology*, vol. 264, p. 225-234, janv. 2019, doi: 10.1016/j.agrformet.2018.10.013.

[15] L. Quan *et al.*, « Maize seedling detection under different growth stages and complex field environments based on an improved Faster R–CNN », *Biosystems Engineering*, vol. 184, p. 1-23, août 2019, doi: 10.1016/j.biosystemseng.2019.05.002.

[16] J. Ribera, Y. Chen, C. Boomsma, et E. J. Delp, « Counting plants using deep learning », in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, nov. 2017, p. 1344-1348. doi: 10.1109/GlobalSIP.2017.8309180.

[17] H. Xiong, Z. Cao, H. Lu, S. Madec, L. Liu, et C. Shen, « TasselNetv2: in-field counting of wheat spikes with context-augmented local regression networks », *Plant Methods*, vol. 15, nº 1, p. 150, déc. 2019, doi: 10.1186/s13007-019-0537-2.

[18] J. Valente, B. Sari, L. Kooistra, H. Kramer, et S. Mücher, « Automated crop plant counting from very high-resolution aerial imagery », *Precision Agriculture*, vol. 21, nº 6, p. 1366-1384, déc. 2020, doi: 10.1007/s11119-020-09725-3.

[19] Y. Liu, C. Cen, Y. Che, R. Ke, Y. Ma, et Y. Ma, « Detection of Maize Tassels from UAV RGB Imagery with Faster R-CNN », *Remote Sensing*, vol. 12, nº 2, 2020, doi: 10.3390/rs12020338.

[20] Z. Lin et W. Guo, « Sorghum Panicle Detection and Counting Using Unmanned Aerial System Images and Deep Learning », *Frontiers in Plant Science*, vol. 11, p. 1346, 2020, doi: 10.3389/fpls.2020.534853.

[21] S. Ren, K. He, R. Girshick, et J. Sun, « Faster r-cnn: Towards real-time object detection with region proposal networks », in *Advances in neural information processing systems*, 2015, p. 91-99.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, et L. Fei-Fei, « Imagenet: A large-scale hierarchical image database », in *2009 IEEE conference on computer vision and pattern recognition*, 2009, p. 248-255.

[23] T.-Y. Lin *et al.*, « Microsoft coco: Common objects in context », in *European conference on computer vision*, 2014, p. 740-755.

[24] R. Geirhos *et al.*, « Shortcut Learning in Deep Neural Networks », *arXiv preprint arXiv:2004.07780*, 2020.

[25] W. Lin, K. Hasenstab, G. Moura Cunha, et A. Schwartzman, « Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment », *Scientific Reports*, vol. 10, nº 1, p. 20336, nov. 2020, doi: 10.1038/s41598-020-77264-y.

[26] X. Jin, S. Liu, F. Baret, M. Hemerlé, et A. Comar, « Estimates of plant density of wheat crops at emergence from very low altitude UAV imagery », *Remote Sensing of Environment*, vol. 198, p. 105-114, sept. 2017, doi: 10.1016/j.rse.2017.06.007.

[27] T. Duan *et al.*, « Comparison of ground cover estimates from experiment plots in cotton, sorghum and sugarcane based on images and ortho-mosaics captured by UAV », *Functional Plant Biol.*, vol. 44, nº 1, p. 169-183, nov. 2016, doi: 10.1071/FP16123.

[28] J. Brooks, *COCO Annotator*. 2019. [En ligne]. Disponible sur: https://github.com/jsbroks/coco-annotator/

[29] R. Bansal, G. Raj, et T. Choudhury, « Blur image detection using Laplacian operator and Open-CV », in *2016 International Conference System Modeling Advancement in Research Trends (SMART)*, 2016, p. 63-67. doi: 10.1109/SYSMART.2016.7894491.

[30] G. E. Meyer et J. C. Neto, « Verification of color vegetation indices for automated crop imaging applications », *Computers and Electronics in Agriculture*, vol. 63, nº 2, p. 282-293, oct. 2008, doi: 10.1016/j.compag.2008.03.009.

[31] N. Otsu, « A threshold selection method from gray-level histograms [J] », *Automatica*, vol. 11, nº 285-296, p. 23-27, 1975.

[32] P. V. Hough, *Method and means for recognizing complex patterns*. Google Patents, 1962.

[33] R. B. Girshick, J. Donahue, T. Darrell, et J. Malik, « Rich feature hierarchies for accurate object detection and semantic segmentation », *CoRR*, vol. abs/1311.2524, 2013, [En ligne]. Disponible sur: http://arxiv.org/abs/1311.2524

[34] W. Liu *et al.*, « Ssd: Single shot multibox detector », in *European conference on computer vision*, 2016, p. 21-37.

[35] K. Chen *et al.*, « MMDetection: Open MMLab Detection Toolbox and Benchmark », *arXiv preprint arXiv:1906.07155*, 2019.

[36] A. Paszke *et al.*, « PyTorch: An Imperative Style, High-Performance Deep Learning Library », in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, et R. Garnett, Éd. Curran Associates, Inc., 2019, p. 8024-8035. [En ligne]. Disponible sur: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[37] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, et S. Belongie, *Feature Pyramid Networks for Object Detection*. 2017.

[38] K. He, X. Zhang, S. Ren, et J. Sun, *Deep Residual Learning for Image Recognition*. 2015.

[39] S. Ghosal *et al.*, « A weakly supervised deep learning framework for sorghum head detection and counting », *Plant Phenomics*, vol. 2019, p. 1525874, 2019.

[40] D. Kuznichov, A. Zvirin, Y. Honen, et R. Kimmel, « Data Augmentation for Leaf Segmentation and Counting Tasks in Rosette Plants », in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, juin 2019, p. 2580-2589. doi: 10.1109/CVPRW.2019.00314.

8 Supplementary material (to put in an external file for submission)

Crop	Maize	Sugarbeet	Sunflower
Early (1)	**12**	**14-15**	**14-16 or germination not over**
Intermediate	**13**	**16**	**17-18**
Late	**14-15**	**17-19**	**19**

Table S1. Correspondance between the "Early", "Intermediate" and "Late stage" and the BBCH scale for each crop

Rules	Parameter name	Operations	Definition	values
Get BW mask	*Excess Green threshold*	Segmentation	The threshold used to transform the image into a vegetation mask	Determined by the otsu method
Find row	*Row number spacing*	Row detection	Expected number of rows	Determined in Table 1

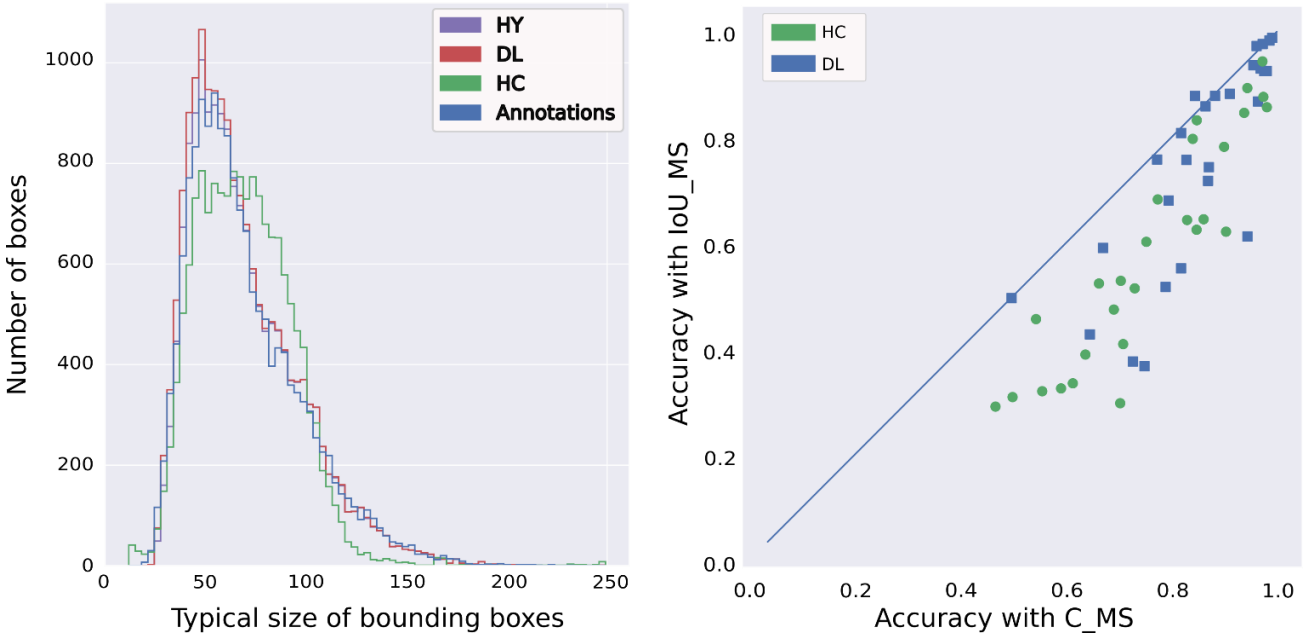
	Row spacing prior	Row detection	Prior value of the row spacing as defined in Table 1	Determined in Table 1
	Peak prior	Row detection	The fraction of the maximum height of the peaks used to consider a peak as corresponding to a row	0.5
	Plant spacing prior	Split object	Prior value of the plant spacing as defined in Table 1	Determined in Table 1
Find plant	*Minimum distance to row*	Weed elimination	Minimum distance from the row centre (expressed relatively to *Row spacing prior*) to consider the objects as weeds	0.25
Remove false positives	*Big Plants Tolerance*	Leaves detection	All centroids under *Big Plants Tolerance* x *Plant spacing prior* are considered to belong to the same plant	0.9

Table S2. List of parameters used for row extraction and plant identification

Figure S3: Justification of a centroid matching strategy Centroid matching strategy (C_MS) is preferred to the IoU one (IoU_MS)

The C_MS was initially compared with an intersection over union matching strategy (IoU_MS), which is more commin The IoU_MS is based on the Intersection over Union between the detected and labelled BB with a standard threshold of 0.5. A detected plant is considered true positive (TP) if its IoU is larger than 0.5. Otherwise, it is a false positive (FP). If a labeled BB has no overlap with any detected BB, it is classified as false negative (FN).

The size of BB of plants detected by the HC method have different dimensions as compared to the labelled BB (Figure 4, left): The distribution of the size of BB for HC is gaussian, while that of labelled, DL and HY are very similar and skewed with significantly smaller BBs as well as larger ones. That means that the HC is missing small object with the IoU_MS. This resulted into lower values of accuracy computed with IoU_MS (Figure 4, right) because of a significant amount of mismatch between the predicted and reference BBs at IoU=0.5. Rather than adapting the IoU threshold level, the distance between centroids is preferred to evaluate the match between predicted and interactively labeled plants. The accuracy computed with C_MS (Figure 4, right) is significantly larger than that computed with IoU_MS, particularly for the low accuracy values as well as for the HC method for the reasons exposed above. Therefore, in the following, the centroid distance is used to compute the terms of the confusion matrix and the accuracy. Detailed metrics can be found in Table S2.



631

632 **Figure S3: Left: distribution of the typical size of BB annotated and those defined around the**
633 **plants identified by the HC method. Right: comparison of Accuracy computed either with**
634 **IoU_MS, and with C_MS for HC (green discs), and DL methods (blue squares).**

635

636 **Table S4. Complete results for the three methods on all sessions. Accuracy, precision and recall**
637 **are presented with the IoU matching strategy.**

638

639

640

641

2.3 Conclusion

Chapter 1 demonstrated that with the correct training information, DL is generally more accurate than handcrafted methods. However, the possible domain shift prevents to use of DL for breeding applications without any human supervision. The hybrid method can slightly improve the robustness capacity of DL methods, but additional data is essential. While the diversity of the proposed dataset was high, the still limited size of the training and testing datasets prevents us from drawing firm conclusions on the factors that degrade performances, including image blurriness, soil type or the presence of artefacts such as rocks. Our dataset is available on Zenodo https://zenodo.org/record/4890370#.YTm7sBnityw, and we hope it can help the community to benchmark their solutions.

3 Design of a large and diverse dataset for training and evaluating deep learning models: application to wheat head detection

3.1 Foreword

Chapter 1 demonstrated the need for a large and diverse dataset to analyze the robustness of the Deep Learning method. Building such datasets for plant phenotyping is rare and mainly dedicated to the controlled conditions installations. For field conditions, it is almost inexistent because up to very recently, acquiring data from HTPP platforms was not as mature as today. Further, the limited labelling capacity of research institutions prevented to build of large datasets. Chapter 2 explores how to re-use the datasets from several sources with different acquisition and labelling protocols. It aims at harmonizing the datasets by applying specific processing of the images and labels. The resulting dataset can then be used to develop robust models to detect wheat heads. Head detection in wheat is essential for estimating the head density and exploring their spatial distribution and the mandatory pre-processing step before characterizing wheat heads. Wheat is also a crop well studied globally, making it possible to cover use cases across contrasted situations globally.

Labelled wheat head datasets were already existing, as the one compiled by Madec et al. [1]. However, they were not always readily available on the internet and were not always well documented. The challenge of chapter 2 is to build a large dataset used to train and evaluate models for head detection. For this purpose, we propose solutions for the labelling issue, harmonising the observations and documenting the diversity of the dataset and the possible confounding factors. This dataset will be later used for crowdsourcing models through open challenges. This chapter is made of two successive papers. The first one presents the dataset used for the first wheat head detection challenge. Based on the results of this first challenge, the original dataset was revised, extended and re-organized to serve a second wheat head detection challenge. The second paper describes this improved Global Whead Head Detection dataset.

3.2 Global Wheat Head Detection 2020

AAAS
Plant Phenomics
Volume 2020, Article ID 3521852, 12 pages
https://doi.org/10.34133/2020/3521852

Plant Phenomics
A SCIENCE PARTNER JOURNAL

Research Article

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,[1,2] **Simon Madec**,[1,2] **Pouria Sadeghi-Tehran**,[3] **Helge Aasen**,[4] **Bangyou Zheng**,[5] **Shouyang Liu**,[2,6] **Norbert Kirchgessner**,[4] **Goro Ishikawa**,[7] **Koichi Nagasawa**,[8] **Minhajul A. Badhon**,[9] **Curtis Pozniak**,[10] **Benoit de Solan**,[1] **Andreas Hund**,[4] **Scott C. Chapman**,[5,11] **Frédéric Baret**,[2,6] **Ian Stavness**,[9] **and Wei 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, UK*
[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, Nishitokyo City, Tokyo, Japan*

Correspondence should be addressed to Ian Stavness; ian.stavness@usask.ca and Wei Guo; guowei@isas.a.u-tokyo.ac.jp

Received 25 April 2020; Accepted 1 July 2020; Published 20 August 2020

The detection of wheat heads in plant images is an important task for estimating pertinent wheat traits including head population density and head characteristics such as health, size, maturity stage, and the presence of awns. Several studies have developed methods for wheat head detection from high-resolution RGB imagery based on machine learning algorithms. However, these methods have generally been calibrated and validated on limited datasets. High variability in observational conditions, genotypic differences, development stages, and head orientation makes wheat head detection a challenge for 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 of wheat images, called the Global Wheat Head Detection (GWHD) dataset. It contains 4700 high-resolution RGB images and 190000 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 dataset is publicly available at http://www.global-wheat.com/and aimed at developing and benchmarking methods for wheat head detection.

1. Introduction

Wheat is the most cultivated cereal crop in the world, along with rice and maize. Wheat breeding progress in the 1950s was vital for food security of emerging countries when Norman Borlaug developed semidwarf kinds of wheat 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. Innovations that increase genetic gain may come from genomic selection, new high-throughput phenotyping techniques, or a combination of both [4–7]. These techniques are essential 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 density (the number of wheat heads 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 of around 10% [8, 9]. Thus, developing image-based methods to increase the throughput and accuracy of counting wheat heads in the field is needed to help breeders manipulate the balance between yield components (plant number, head density, grains per head, grain weight) in their breeding selections.

Thanks to increases in GPU performance and the emergence of large-scale datasets [10, 11], deep learning has become the state of the art approach for many computer vision tasks, including 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 high-resolution RGB 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. Gibbs et al. [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, colour, 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, and wheat heads often overlap and occlude each other in fields with higher planting densities.

A common strategy for handling limited datasets is to train a CNN model on a portion of a phenotyping trial field and test it on the remaining portion of the field [25]. This is a fundamental flaw of empirical approaches against causal models: there is no theoretical guarantee that a CNN model is robust on new acquisitions. In addition, a comparison between methods from different authors requires 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 are processed. Finally, labelling protocols may be different between institutions, which will limit model performance 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, and 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 comprising the GWHD dataset 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 30.5 cm (USask_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 GWHD dataset covers a range of pedoclimatic conditions including very productive context such as the loamy soil of the 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, and NAU_1, the experiments were designed to compare irrigated and water-stressed environments.

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

Sub-dataset name	Institution	Country	Lat (°)	Long (°)	Year	No. of dates	Targeted stages	Row spacing (cm)	Sowing density (seeds·m^2)	No. of genotypes
UTokyo_1	NARO & UTokyo	Japan	36.0 N	140.0 E	2018	3	Postflowering	15	186	66
UTokyo_2	NARO & UTokyo	Japan	42.8 N	143.0 E	2016	6	Flowering*	12.5	200	1
Arvalis_1	Arvalis	France	43.7 N	5.8 E	2017	3	Postflowering-ripening	17.5	300	20
Arvalis_2	Arvalis	France	43.7 N	5.8 E	2019	1	Postflowering	17.5	300	20
Arvalis_3	Arvalis	France	49.7 N	3.0 E	2019	3	Postflowering-ripening	17.5	300	4
INRAE_1	INRAE	France	43.5 N	1.5 E	2019	1	Postflowering	16	300	7
USask_1	University of Saskatchewan	Canada	52.1 N	106 W	2019	1	n.a	30.5	250	16
RRes_1	Rothamsted research	UK	51.8 N	0.36 W	2016	1	n.a	n.a	350	6
ETHZ_1	ETHZ	Switzerland	47.4 N	8.6 E	2018	1	n.a	12.5	400	354
NAU_1	Nanjing Agric. University	China	31.6 N	119.4 E	2018	1	Flowering*	20	300 or 450	5
UQ_1	UQueensland	Australia	27.5 S	152.3 E	2016	1	Flowering -ripening	22	150	8

*Images were checked carefully to ensure that heads have fully developed and flowered.

Table 2: Image characteristics of the sub-datasets comprising the GWHD dataset. All cameras looked vertically downward.

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	800 × 800†	1.8	0.56
Arvalis_3	Handheld	Sony RX0	7.7	9.99	Automatic	800 × 800†	1.8	0.56
INRAE_1	Handheld	Sony RX0	7.7	9.99	Automatic	800 × 800†	1.8	0.56
USask_1	Minivehicle	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

*The field of view is measured diagonally. The reported measure is the half-angle. †Original images were cropped, and a subimage of size 800 × 800 was extracted from the central area. §The camera was positioned perpendicular to the ground and automatically adjusted to ensure a 2.2 m distance was maintained between the camera and canopy.

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 in diameter, the acquired GSDs are high enough to detect heads and even awns visually. Although all images were acquired at the nadir-viewing direction, some geometric distortion may be observed for a few sub-datasets due to the different lens characteristics of the cameras used.

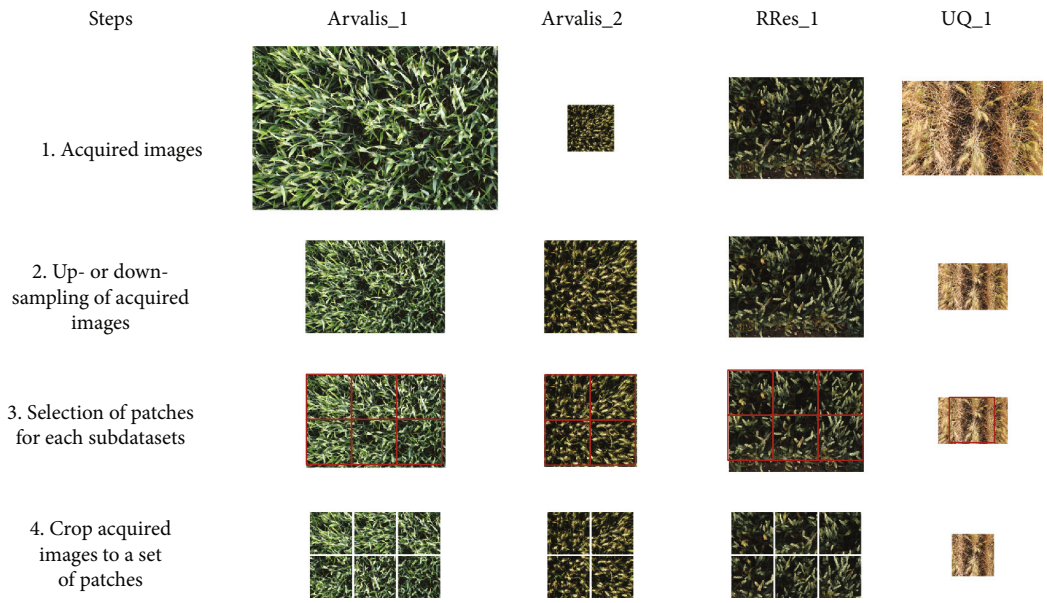


Figure 1: Overview of the harmonization process conducted. Images were first rescaled using bilinear interpolation up- or downsampling techniques. Then, the rescaled images were split into 1024×1024 squared patches.

Datasets UTokyo_1 and ETHZ_1 are particularly affected by this issue. Each institution acquired images from different platforms, including handheld, cart, minivehicle, and gantry systems. The diversity of camera sensors and acquisition configurations resulted in a wide range of image properties, which will assist in training deep learning models to better generalize across different image acquisition conditions.

2.3. Data Harmonization. An important aspect of assembling the GWHD dataset was harmonizing the various sub-datasets (Figure 1). A manual inspection of images was first conducted to ensure that they could be well interpreted. Images acquired at too early of a growth stage were removed when heads were not clearly visible (Figure 2(d)). Most of the images were also acquired before the appearance of head senescence since heads tend to overlap when the stems start to bend at this stage.

Object scale, i.e., the size of the object in pixels, is an important factor in the design of object 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, but are generally around 1.5 cm in diameter and 10 cm in length. The actual image resolution, at the level of wheat heads, varied significantly between sub-datasets: the GSD varies by a factor of 5 (Table 1) while the actual resolution at the head level also depends 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 downsample the original images. The scaling factor applied to each sub-dataset is displayed in Table 3.

Most deep learning algorithms are trained with square-sized image patches. When the original images were cropped into square patches, the size of the patches was selected to reduce the chance that heads would cross the edges of the patches and be partly cut off. Images were therefore split into 1024×1024 squared patches containing roughly 20 to 60 heads each, with only a few heads crossing the patch edges. The number of patches per original image varied from 1 to 6 depending on the sub-dataset (Table 3). These squared patches will be termed "images" for the remainder of the paper.

2.4. Labelling. A web-based labelling platform was developed 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)).

Several institutions had 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

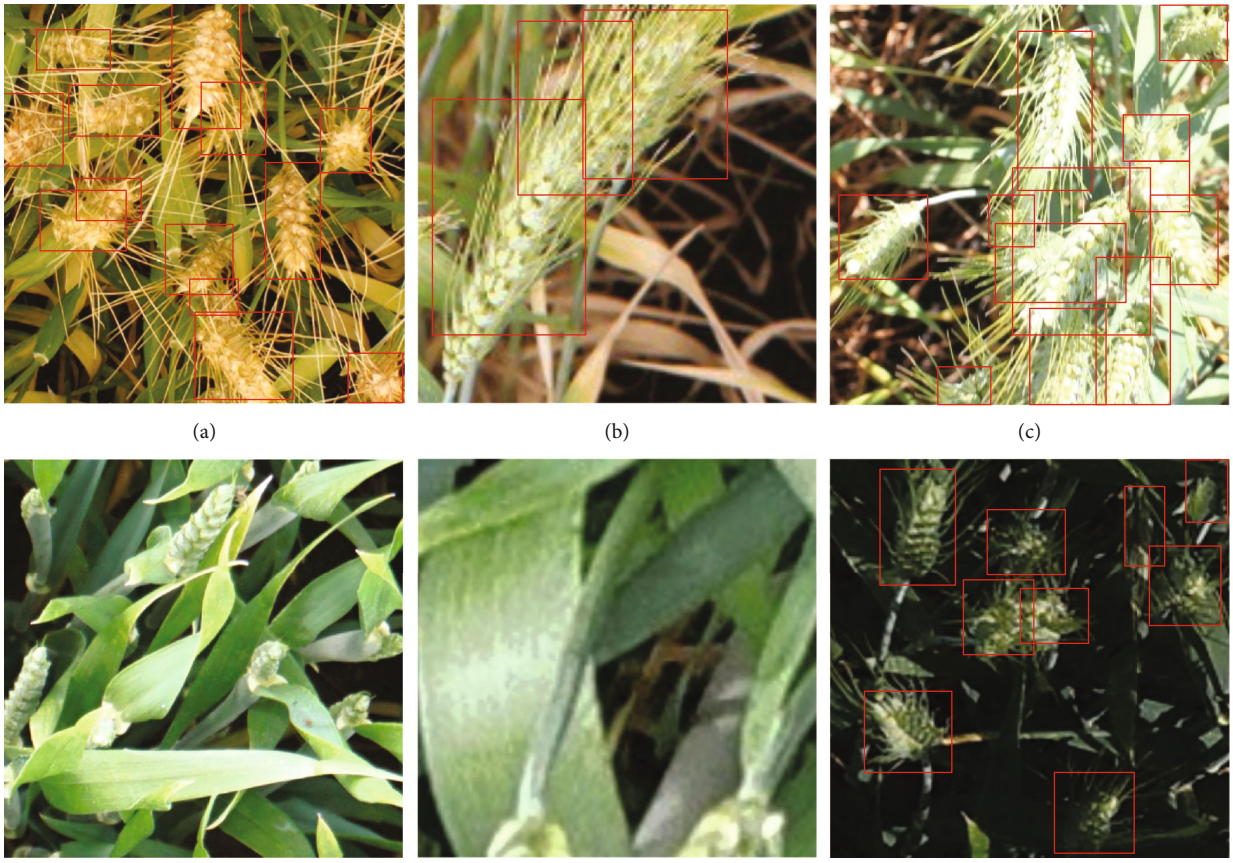


Figure 2: Examples of wheat heads difficult to label. These examples are zoomed-in views from images contained in the dataset, with different zoom factors. It includes overlapping heads (a–c), heads at emergence (d), heads that are partly cut at the border of the image (e), and images with a low illumination (f). Note that image (d) was removed from the dataset because of the ambiguity of heads at emergence. Wheat heads in the image (e) were not labelled because less than 30% of their basal part is visible, as defined in Section 2.4.

Table 3: Statistics for each component of the Global Wheat Head Detection.

Sub-dataset name	No. of acquired images	No. of patch per image	Original GSD (mm)	Sampling factor	Used GSD (mm)	No. of labelled images	No. of labelled heads	Average no. 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	—

*Some labelled images have been removed during the labelling process.

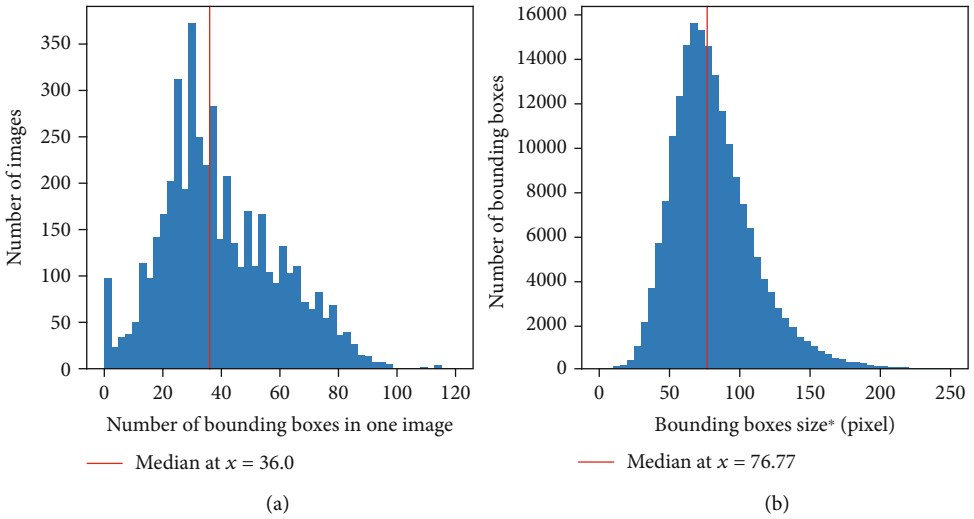


Figure 3: Distribution of the number of bounding boxes per image (a) and bounding boxes size* (b) in the GWHD dataset. *The bounding box size is defined as the square root of the bounding box area in pixel.

trained over UTokyo_1 and Arvalis_1 sub-datasets and then applied to the unlabelled sub-datasets. Boxes with an associated confidence score greater than and equal to 0.5 were retained and proposed to the user for correction. This semi-automatic active learning increased the throughput of the labelling process by a factor of four as compared to a fully manual process. The process is detailed in Figure S1.

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

3. Description of the Dataset

3.1. General Statistics. The GWHD dataset represents 4698 squared patches extracted from the 2219 original high-resolution RGB images acquired across the 11 sub-datasets (Table 3). It represents 188445 labelled heads which average 40 heads per image in good agreement with the 20 to 60 targeted heads per image. However, the distribution among and within sub-datasets is relatively broad (Figure 3(a)). We included about 100 images that contain no heads to represent in-field capturing conditions and add difficulty for benchmarking. Few images contain more than 100 heads with a maximum of 120 heads. Multiple peaks corresponding to the several sub-datasets (Figure 3(b)) can be observed due 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 3(b)). The typical dimension is computed as the square root of the area. It corresponds well to the targeted scale, i.e., 1.5 cm × 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 correspond exactly to the viewing geometry of the RGB cameras. The harmonization of object scale across sub-datasets can be further confirmed visually in Figure 4.

3.2. Diversity of Sampled Genotypes, Environments, and Developmental Stages. The diversity of acquisition conditions sampled by the GWHD dataset is well illustrated in Figure 4: illumination conditions are variable, with a wide range of heads and background appearance. Further, we observe variability in head orientation and view directions, from an almost nadir direction up to a mostly oblique direction as in the case of ETHZ_1 (Figure 4). A selection of bounding boxes extracted from the several sub-datasets (Figure 5) shows a 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 colour, and blurriness. In addition, a few heads were cut off when the bounding box crossed the edge of the image.

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 instance 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 [17]. 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 [31, 32]. The GrassClover dataset includes images of forage fields and semantic segmentation labels for grass, clover, and weed vegetation types [33]. Datasets for counting sorghum [27, 34] and wheat heads [35] have also been published with dot annotations.

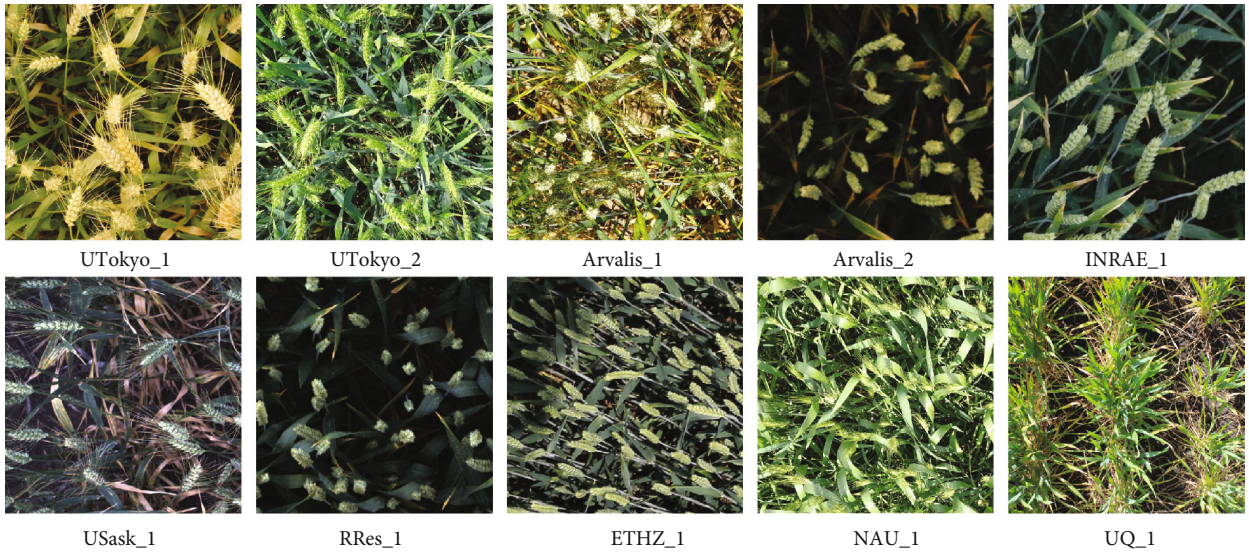


Figure 4: Example of images from different acquisition sites after cropping and rescaling.

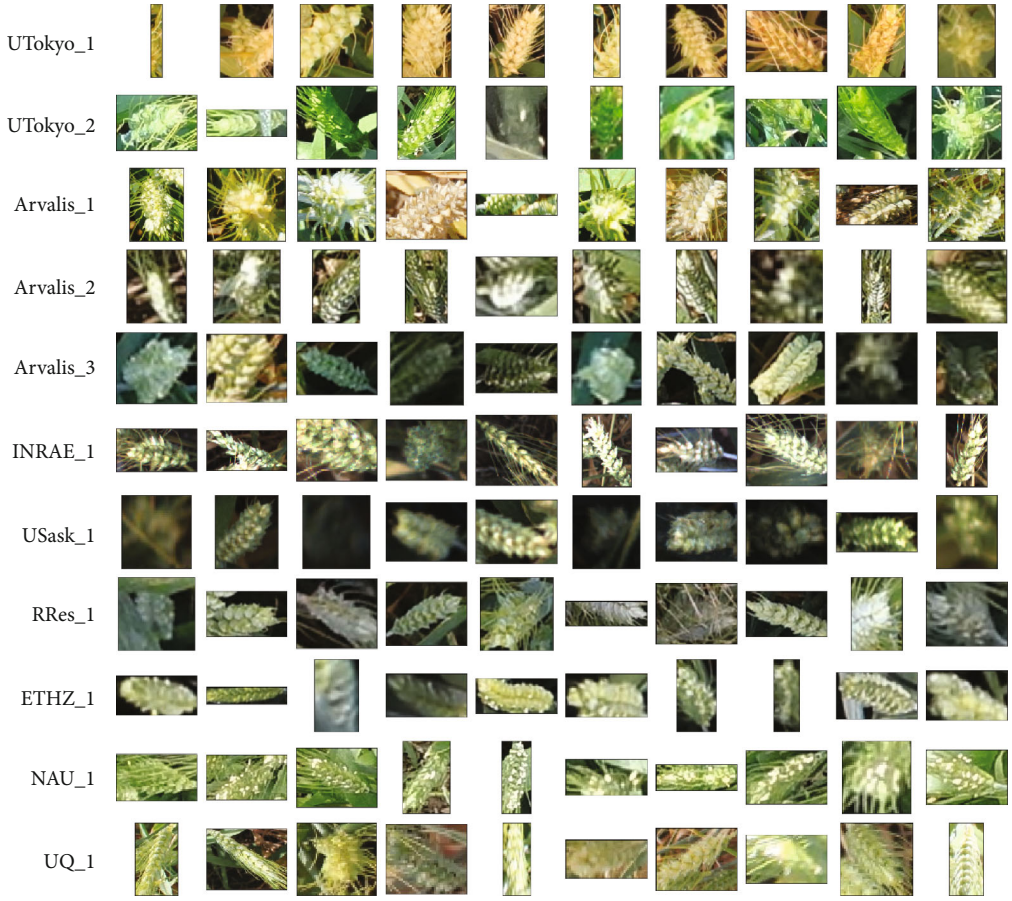


Figure 5: A selection of bounding boxes for each sub-dataset. The same size of pixels is used across all the bounding boxes displayed.

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 [36] 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 [37] or Open Images V4 [38] 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).

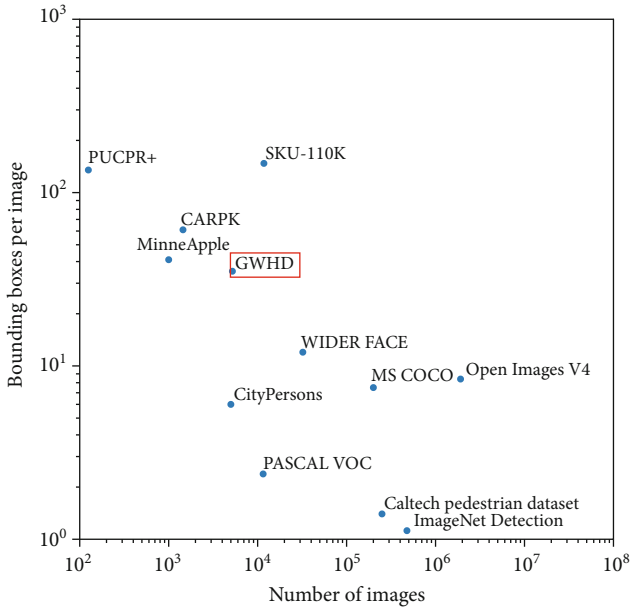


Figure 6: Comparison of GWHD dataset with other object detection datasets. Both axes are in log-scale.

However, some specific datasets like PUCPR [39], CARPK [40], and SKU-110K [41] are tailored to the problem of detecting objects (e.g., cars, products) in dense contexts. They have a much higher object density than the 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 that have a wide variability of appearance as demonstrated previously, surrounded by a very diverse background which would constitute a more difficult problem than detecting cars or densely packed products on store shelves.

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 high-resolution RGB images. An open machine learning competition will be held from May to August 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. The test dataset includes all the images from Australia, Japan, and China, representing 1276 images to evaluate model performance, including robustness against unseen images.

4.2. Evaluation Metrics. The choice of bounding boxes as labels in the GWHD dataset allows it to be used for object detection. The mean average precision computed from the true and false positives is usually used to quantify performance in object detection tasks. A true positive corresponds to a predicted bounding box with an intersection over union (IoU) greater than and equal to 0.5 with the closest labelled bounding box. A false positive corresponds to a predicted bounding box with an IoU strictly lower than 0.5 with the closest labelled bounding box. In the case of two predicted boxes with an IoU greater than or equal to 0.5 on the same bounding box, the most confident one is considered as a true positive and the other as a false positive. The mean Average Precision noted as mAP@0.5 is the considered metric for evaluating the localization performance. Detection of individual wheat heads is required for characterizing their size, inclination, colour, or health. 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 as the Root Mean Square Error (RMSE), relative RMSE (rRMSE), and Coefficient of Determination (R^2) to quantify the performance of object 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 [12], with a ResNet34 and ResNet50 as the backbone. Faster-RCNN is one of the most popular object detection models and used in Madec et al. [8]. ResNet34 is used along with ResNet50 because it is less prone to overfitting and faster to train. Due to memory constraints, the input size was set to 512×512 pixels. We randomly sampled ten patches of size 512×512 pixels for each image in the training dataset resulting in a training dataset composed of 34220 patches. We predicted on a set of overlapping patches of size 512×512 pixels regularly extracted from the test images of size 1024×1024 pixels and then merged the results. After 10 epochs, representing 342200 iterations in total, the best model is obtained at epoch 3 for both backbones. It yielded a mAP@0.5 of 0.77 and a mean RMSE of 12.63 wheat heads per image which corresponds to rRMSE = 39%. The coefficient of regression is 0.57. All results are provided in Figure S2. The relatively poor performance of a standard object detection network on the GWHD dataset provides an opportunity for substantial future improvement with novel methods. The GWHD competition is expected to instigate new wheat head detection approaches that will provide more accurate results.

Table 4: The minimum metadata that should be associated with images of wheat heads.

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

[*]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. [‡]The distance between camera and canopy is an essential piece of information to harmonize dataset and calculate the density and should be carefully monitored during an acquisition. [§]In case of multiple images over the same microplot.

5. Discussion

5.1. Image Acquisition Recommendations. To successfully detect wheat heads, they should be fully emerged and clearly visible within the images, with minimum overlap among heads and leaves. For some genotypes and environmental conditions, we observed that the wheat stems tend to bend for the latest grain filling stages, which increases 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. Therefore, we recommend acquiring images immediately after flowering when the wheat heads have fully emerged and are still upright in the field.

For image acquisition, a near nadir viewing direction is recommended to limit the overlap between heads, especially in the case of populations with high head density. Likewise, a narrow field of view is preferred. However, a narrow field of view may result in a small image footprint when the camera is positioned at a height close to the top of the canopy. Therefore, we recommend increasing the camera height to get a larger sampled area and reduce the number of heads that will be cropped at the edge of the image. The size of the sampled area is important when head identification is used for estimating the head population density. The minimum sampled area should be that of our squared patch, i.e., 1024 × 1024 pixels of 0.4 mm/pixel which corresponds to an area of about 40 cm^2. To achieve this sampled area, while maintaining a narrow field of view of ±15°, the distance between the camera and the top of the canopy should be around 1.0 m. However, a larger sampling area is preferable for head population density estimation, where at least 100 cm^2 should be sampled to account for possible heterogeneity across rows. This would be achieved with a 2.5 m distance between the camera and the top of the canopy.

When estimating wheat head density, i.e., the number of heads per unit ground area, accurate knowledge of the sampled area is critical. The nonparallel geometry of image acquisition, with significative "fisheye" lens distortion effects, induces uncertainty about the sampled area. Even for our typical case with limited distortion effects (±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 measurement. Further, the definition of the reference height at which to compute the sampled area is still an open question, because within a population of wheat plants, the heights of the heads can vary by more than 25 cm, which induces a 21% difference in the sampled area between the lowest and highest head. Further work should investigate this important question.

Finally, our experience suggests that using a submillimetre resolution at the top of the canopy is required for efficient head detection. However, the optimal resolution is yet to be defined. Previous work [8] recommended 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 of wheat imaging, particularly regarding the possibility to use UAV observations for head density estimation in large wheat breeding experiments.

5.2. Minimum Information Associated with the Sub-datasets and FAIR Principles. The FAIR principles (Findable, Accessible, Interoperable, and Reusable [42]) should be applied to the images that populate the GWHD dataset. A minimum set of metadata should be associated with each image as proposed in [43] to verify the FAIR principles. The lack of metadata was an issue for precise data harmonization and is limiting factor for further data interpretation [44] and possible meta-analysis. Therefore, we 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 camera and acquisition vehicle used. Both can be defined at the session level and the image level. Our recommendations are summarized in Table 4. We encourage attaching more metadata such as camera settings (model, white balance correction, et al.) when possible because it adds context for further data reuse.

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 by the current dataset: more than 68% of the images within the GWHD dataset come from Europe and 43% from France. Further, some regions are currently missing, including Africa, Latin America, and the Middle East. As future work, we hope to expand the GWHD dataset in order to get a more comprehensive dataset. Therefore, we 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 encourage potential contributors to contact the corresponding authors through http://www.global-wheat.com. We also plan to extend the GWHD dataset in the future for classification and segmentation tasks at the wheat head level, for instance, the size of the wheat head or flowering state. This expansion would require an update of the current labels.

6. Conclusion

Object detection methods for localizing and identifying wheat heads in images are useful for estimating head density in wheat populations. Head detection may also be considered as a first step in the search for additional wheat traits, including the spatial distribution between rows, the presence of awns, size, inclination, colour, grain filling stage, and health. These traits may prove useful for wheat breeders and some may help farmers to better manage their crops.

In order to improve the accuracy and reliability of wheat head detection and localization, we have assembled the Global Wheat Head Detection dataset—an extensive and diverse dataset of wheat head images. It is designed to develop and benchmark head detection methods proposed by the community. It represents a large collaborative international effort. An important contribution 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 wheat heads. We hope that these guidelines will enable practitioners to expand the GWHD dataset in the future with additional sub-datasets that represent even more genotypic and environmental diversity. The GWHD dataset has been proposed together with an open research competition to find more accurate and robust methods for wheat head detection across the wide range of wheat growing regions around the world. The solutions proposed in the competition will be made open-source and shared with the plant phenotyping community.

Conflicts of Interest

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

Authors' Contributions

E.D., S.M., B.S, and F.B. organized the field experiment and data collection for France dataset. P.S.T. organized the field experiment and data collection for U.K. dataset. H.A., N.K., and A.H. organized the field experiment and data collection for Switzerland dataset. G.I., K.N., and W.G. organized the field experiment and data collection for Japan dataset. S.L. and F.B. organized the field experiment and data collection for China dataset. C.P., M.B., and I.S. organized the field experiment and data collection for Canada dataset. B.Z. and S.C.C. organized the 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 prelabelling 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.

Acknowledgments

The French team received support from ANRT for the CIFRE grant of Etienne David, cofunded by Arvalis. The study was partly supported by several projects including ANR PHENOME, ANR BREEDWHEAT, CASDAR LITERAL, and FSOV "Plastix". Many thanks are due to the people who annotated the French datasets, including Frederic Venault, Xiuliang Jin, Mario Serouard, Ilias Sarbout, Carole Gigot, Eloïse Issert, and Elise Lepage. The Japanese team received support from JST CREST (Grant Numbers JPMJCR16O3, JPMJCR16O2, and JPMJCR1512) and MAFF Smart-Breeding System for Innovative Agriculture (BAC1003), Japan. Many thanks are due to the people who annotated the Japanese dataset, including Kozue Wada, Masanori Ishii, Ryuuichi Kanzaki, Sayoko Ishibashi, and 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 are due 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 are due to Jie Zhou and many volunteers from Nanjing Agricultural University to accomplish the annotation. The dataset collection at ETHZ was supported by Prof. Achim Walter, who leads the Crop Science group. Many thanks are due 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 the annotation of the ETHZ dataset and to Brigita Herzog and Hansueli Zellweger for crop husbandry.

Supplementary Materials

Figure S1: the proposed "weakly supervised deep learning framework" to prelabel images efficiently. Figure S2: epoch-wise results (RMSE, rRMSE, R2, mAP@0.5) of Faster-RCNN baseline with ResNet34 and ResNet50. The best model is obtained at epoch 3 for both backbones. *(Supplementary Materials)*

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, pp. 95–110, 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," *Field Crops Research*, vol. 119, no. 1, pp. 201–212, 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," *Scientific Reports*, vol. 8, no. 1, article 16865, 2018.

[4] M. Reynolds, S. Chapman, L. Crespo-Herrera et al., "Breeder friendly phenotyping," *Plant Science*, vol. 295, article 110396, 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, 2018.

[6] A. Hund, L. Kronenberg, J. Anderegg, K. Yu, and A. Walter, "Non-invasive phenotyping of cereal growth and development characteristics in the field," in *Advances in Crop Breeding Techniques*, Burleigh Dodds, 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, 2015.

[8] S. Madec, X. Jin, H. Lu et al., "Ear density estimation from high resolution RGB imagery using deep learning technique," *Agricultural and Forest Meteorology*, vol. 264, pp. 225–234, 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*, vol. 14, no. 1, article 100, 2018.

[10] M. Z. Alom, T. M. Taha, C. Yakopcic et al., "A state-of-the-art survey on deep learning theory and architectures," *Electron*, vol. 8, no. 3, p. 292, 2019.

[11] O. Russakovsky, J. Deng, H. Su et al., "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds., vol. 9351 of Lecture Notes in Computer Science, pp. 234–241, Springer, Cham, 2015.

[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*, 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," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 2055–2063, Venice, Italy, 2017.

[18] J. R. Ubbens and I. Stavness, "Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks," *Frontiers in Plant Science*, vol. 8, article 1190, 2017.

[19] A. K. Singh, B. Ganapathysubramanian, S. Sarkar, and A. Singh, "Deep learning for plant stress phenotyping: trends and future perspectives," *Trends in Plant Science*, vol. 23, no. 10, pp. 883–898, 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," *Frontiers in Plant Science*, vol. 10, 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 Physiology*, vol. 181, no. 1, pp. 28–42, 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, 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," *Frontiers in Plant Science*, vol. 10, article 1749, 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, article 4820305, pp. 1–10, 2019.

[26] J. Brooks, *COCO Annotator*, 2019.

[27] S. Ghosal, B. Zheng, S. C. Chapman et al., "A weakly supervised deep learning framework for sorghum head detection and counting," *Plant Phenomics*, vol. 2019, article 1525874, pp. 1–14, 2019.

[28] J. Redmon and A. Farhadi, *YOLOv3: An Incremental Improvement*, 2018.

[29] H. Scharr, M. Minervini, A. Fischbach, and S. A. Tsaftaris, "Annotated image datasets of rosette plants," pp. 1–16, 2014.

[30] H. Uchiyama, S. Sakurai, M. Mishima et al., "An easy-to-setup 3D phenotyping platform for KOMATSUNA dataset," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW),*, pp. 2038–2045, Venice, Italy, 2017.

[31] I. Sa, M. Popović, R. Khanna et al., "Weedmap: a large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming," *Remote Sensing*, vol. 10, no. 9, article 1423, 2018.

[32] 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*, vol. 18, no. 5, article 1580, 2018.

[33] S. Skovsen, M. Dyrmann, A. K. Mortensen et al., "The grass clover image dataset for semantic and hierarchical species understanding in agriculture," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, 2019.

[34] W. Guo, B. Zheng, A. B. Potgieter et al., "Aerial imagery analysis–quantifying appearance and number of sorghum heads for applications in breeding and agronomy," *Frontiers in Plant Science*, vol. 9, article 1544, 2018.

[35] 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.

[36] N. Hani, P. Roy, and V. Isler, "MinneApple: a benchmark dataset for apple detection and segmentation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 852–858, 2020.

[37] T. Y. Lin, M. Maire, S. Belongie et al., "Microsoft COCO: common objects in context," in *Computer Vision – ECCV 2014. ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8693 of Lecture Notes in Computer Science, pp. 740–755, Springer, Cham, 2014.

[38] A. Kuznetsova, H. Rom, N. Alldrin et al., "The Open Images Dataset V4: unified image classification, object detection, and visual relationship detection at scale," *International Journal of Computer Vision*, vol. 128, pp. 1956–1981, 2020.

[39] 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 Systems with Applications*, vol. 42, no. 11, pp. 4937–4949, 2015.

[40] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *Oct 2017 in 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4165–4173, Venice, Italy, 2017.

[41] E. Goldman, R. Herzig, A. Eisenschtat, J. Goldberger, and T. Hassner, "Precise detection in densely packed scenes," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5222–5231, Long Beach, CA, USA, 2019.

[42] C. Pommier, C. Michotey, G. Cornut et al., "Applying FAIR principles to plant phenotypic data management in GnpIS," *Plant Phenomics*, vol. 2019, article 1671403, pp. 1–15, 2019.

[43] H. Ćwiek-Kupczyńska, T. Altmann, D. Arend et al., "Measures for interoperability of phenotypic data: minimum information requirements and formatting," *Plant Methods*, vol. 12, no. 1, article 44, 2016.

[44] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 924–933, Santa Rosa, CA, USA, 2017.

3.3 Global Wheat Head Detection 2021

AAAS
Plant Phenomics
Volume 2021, Article ID 9846158, 9 pages
https://doi.org/10.34133/2021/9846158

Plant Phenomics
A SCIENCE PARTNER JOURNAL

Database/Software Article

Global Wheat Head Detection 2021: An Improved Dataset for Benchmarking Wheat Head Detection Methods

Etienne David,[1,2] **Mario Serouart**,[1,2] **Daniel Smith**,[3] **Simon Madec**,[1,3] **Kaaviya Velumani**,[2,4] **Shouyang Liu**,[5] **Xu Wang**,[6] **Francisco Pinto**,[7] **Shahameh Shafiee**,[8] **Izzat S. A. Tahir**,[9] **Hisashi Tsujimoto**,[10] **Shuhei Nasuda**,[11] **Bangyou Zheng**,[12] **Norbert Kirchgessner**,[13] **Helge Aasen**,[13] **Andreas Hund**,[13] **Pouria Sadhegi-Tehran**,[14] **Koichi Nagasawa**,[15] **Goro Ishikawa**,[16] **Sébastien Dandrifosse**,[17] **Alexis Carlier**,[17] **Benjamin Dumont**,[18] **Benoit Mercatoris**,[17] **Byron Evers**,[6] **Ken Kuroki**,[19] **Haozhou Wang**,[19] **Masanori Ishii**,[19] **Minhajul A. Badhon**,[20] **Curtis Pozniak**,[21] **David Shaner LeBauer**,[22] **Morten Lillemo**,[8] **Jesse Poland**,[6] **Scott Chapman**,[3,12] **Benoit de Solan**,[1] **Frédéric Baret**,[2] **Ian Stavness**,[20] and **Wei Guo**[19]

[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]*School of Food and Agricultural Sciences, The University of Queensland, Gatton, 4343 QLD, Australia*
[4]*Hiphen SAS, 120 Rue Jean Dausset, Agroparc, Bâtiment Technicité, 84140 Avignon, France*
[5]*Plant Phenomics Research Center, Nanjing Agricultural University, Nanjing, China*
[6]*Wheat Genetics Resource Center, Dep. of Plant Pathology, Kansas State Univ., 4024 Throckmorton Plant Sciences Center, Manhattan, Kansas, USA*
[7]*Global Wheat Program, International Maize and Wheat Improvement Centre (CIMMYT), Mexico, D.F., Mexico*
[8]*Faculty of Biosciences, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Ås, Norway*
[9]*Agricultural Research Corporation, Wheat Research Program, P.O. Box 126, Wad Medani, Sudan*
[10]*Arid Land Research Center, Tottori University, Tottori 680-0001, Japan*
[11]*Laboratories of Plant Genetics and Plant Breeding, Graduate School of Agriculture, Kyoto University, Japan*
[12]*CSIRO Agriculture and Food, Queensland Biosciences Precinct, 306 Carmody Road, St Lucia, 4067 QLD, Australia*
[13]*Institute of Agricultural Sciences, ETH Zurich, Universitätstrasse 2, 8092 Zurich, Switzerland*
[14]*Plant Sciences Department, Rothamsted Research, Harpenden, UK*
[15]*Institute of Crop Science, National Agriculture and Food Research Organization, Japan*
[16]*Hokkaido Agricultural Research Center, National Agriculture and Food Research Organization, Japan*
[17]*Biosystems Dynamics and Exchanges, TERRA Teaching and Research Center, Gembloux Agro-Bio Tech, University of Liège, 5030 Gembloux, Belgium*
[18]*Plant Sciences, TERRA Teaching and Research Center, Gembloux Agro-Bio Tech, University of Liège, 5030 Gembloux, Belgium*
[19]*Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Midori-cho, Nishitokyo City, Tokyo, Japan*
[20]*Department of Computer Science, University of Saskatchewan, Canada*
[21]*Department of Plant Sciences, University of Saskatchewan, Canada*
[22]*College of Agriculture and Life Sciences, University of Arizona, Tucson, Arizona, USA*

Correspondence should be addressed to Etienne David; e.david@arvalis.fr

Received 31 May 2021; Accepted 11 August 2021; Published 22 September 2021

The Global Wheat Head Detection (GWHD) dataset was created in 2020 and has assembled 193,634 labelled wheat heads from 4700 RGB images acquired from various acquisition platforms and 7 countries/institutions. With an associated competition hosted in Kaggle, GWHD_2020 has successfully attracted attention from both the computer vision and agricultural science communities. From this first experience, a few avenues for improvements have been identified regarding data size, head diversity, and label reliability. To address these issues, the 2020 dataset has been reexamined, relabeled, and complemented by adding 1722 images from 5 additional countries, allowing for 81,553 additional wheat heads. We now release in 2021 a new version of the Global Wheat Head Detection dataset, which is bigger, more diverse, and less noisy than the GWHD_2020 version.

1. Introduction

Quality training data is essential for the deployment of deep learning (DL) techniques to get a general model that can scale on all the possible cases. Increasing dataset size, diversity, and quality is expected to be more efficient than increasing network complexity and depth [1]. Datasets like ImageNet [2] for classification or MS COCO [3] for instance detection are crucial for researchers to develop and rigorously benchmark new DL methods. Similarly, the importance of getting plant- or crop-specific datasets is recognized within the plant phenotyping community ([4–10], p. 2, [11–13]). These datasets allow benchmarking the algorithm performances used to estimate phenotyping traits while encouraging computer vision experts to further improvement ([10], p. 2, [14–17]). The emergence of affordable RGB cameras and platforms, including UAVs and smartphones, makes in-field image acquisition easily accessible. These high-throughput methods are progressively replacing manual measurement of important traits such as wheat head density. Wheat is a crop grown worldwide, and the number of heads per unit area is one of the main components of yield potential. Creating a robust deep learning model performing over all the situations requires a dataset of images covering a wide range of genotypes, sowing density and pattern, plant state and stage, and acquisition conditions. To answer this need for a large and diverse wheat head dataset with consistent and quality labeling, we developed in 2020 the Global Wheat Head Detection (GWHD_2020) [18] that was used to benchmark methods proposed in the computer vision community and recommend best practices to acquire images and keep track of the metadata.

The GWHD_2020 dataset results from the harmonization of several datasets coming from nine different institutions across seven countries and three continents. There are already 27 publications [19–45] (accessed July 2021) that have reported their wheat head detection model using the GWHD_2020 dataset as the standard for training/testing data. A "Global Wheat Detection" competition hosted by Kaggle was also organized, attracting 2245 teams across the world [14], leading to improvements in wheat head detection models [23, 25, 31, 41]. However, issues with the GWHD_2020 dataset were detected during the competition, including labeling noise and an unbalanced test dataset.

To provide a better benchmark dataset for the community, the GWHD_2021 dataset was organized with the following improvements: (1) the GWHD_2020 dataset was checked again to eliminate few poor-quality images, (2) images were re-labeled to avoid consistency issues, (3) a wider range of developmental stages from the GWHD_2020 sites was included, and (4) datasets from 5 new countries (the USA, Mexico, Republic of Sudan, Norway, and Belgium) were added. The resulting GWHD_2021 dataset contains 275,187 wheat heads from 16 institutions distributed across 12 countries.

2. Materials and Methods

The first version of GWHD_2020, used for the Kaggle competition, was divided into several subdatasets. Each subdataset represented all images from one location, acquired with one sensor while mixing several stages. However, wheat head detection models may be sensitive to the developmental stage and acquisition conditions: at the beginning of head emergence, a part of the head is barely visible because it is still not fully out from the last leaf sheath and possibly masked by the awns. Further, during ripening, wheat heads tend to bend and overlap, leading to more erratic labeling. A redefinition of the subdataset was hence necessary to help investigate the effect of the developmental stage on model performances. The new definition of a subdataset was then formulated as "a consistent set of images acquired over the same experimental unit, during the same acquisition session with the same vector and sensor." A subdataset defines therefore a domain. This new definition forced to split the original GWHD_2020 subdatasets into several smaller ones. The UQ_1 was split into 6 much smaller subdatasets, Arvalis_1 was split into 3 subdatasets, Arvalis_3 into 2 subdatasets, and utokyo_1 into 2 subdatasets. However, in the case of utokyo_2 which was a collection of images taken by farmers at different stages and in different fields, the original subdataset was kept. Overall, the 11 original subdatasets in GWHD_2020 were distributed into 19 subdatasets for GWHD_2021.

Almost 2000 new images were added to GWHD_2020, constituting a major improvement. A part of the new images comes from the institutions already contributing to GWHD_2020 and was collected during a different year and/or at a different location. This was the case for Arvalis (Arvalis_7 to Arvalis_12), University of Queensland (UQ_7 to UQ_11), Nanjing Agricultural University (NAU_2 and NAU_3), and University of Tokyo (Utokyo_1). In addition, 14 new subdatasets were included, coming from 5 new countries: Norway (NMBU), Belgium (Université of Liège [46]), United States of America (Kansas State University [47], TERRA-REF [7]), Mexico (CIMMYT), and Republic of Sudan (Agricultural Research Council). All these images were acquired at a ground sampling distance between 0.2

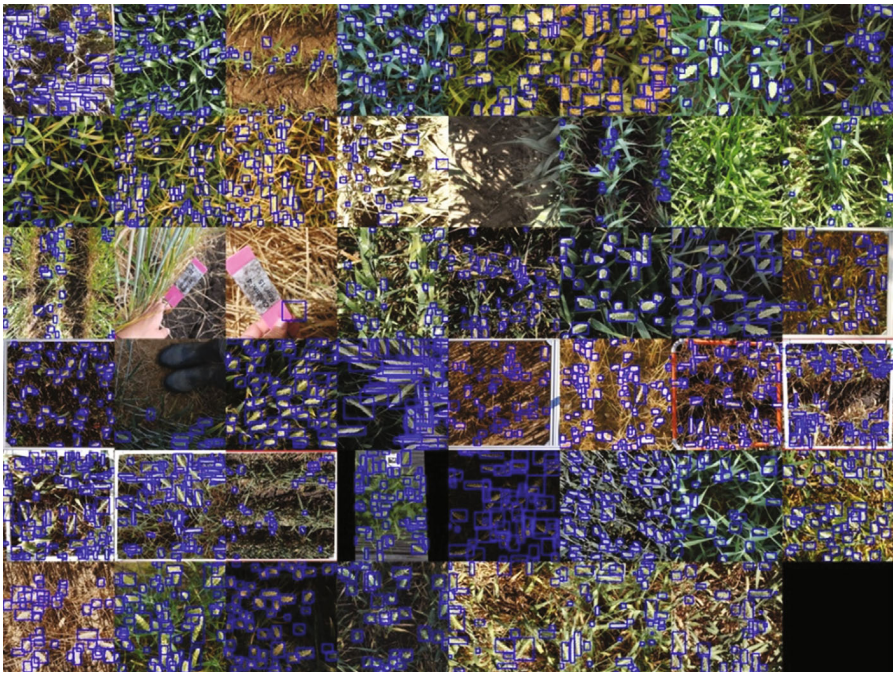


Figure 1: Sample images of the Global Wheat Head Detection 2021. The blue boxes correspond to the interactively labeled heads.

and 0.4 mm, i.e., similar to that of the images in GWHD_2020. Because none of them was already labeled, a sample was selected by taking no more than one image per microplot, which was randomly cropped to 1024×1024 px patches that will be called images in the following for the sake of simplicity.

With the addition of 1722 images and 86,000 wheat heads, the GWHD_2021 dataset contains 6500 images and 275,000 wheat heads. The increase in the number of subdatasets from 18 to 47 leads to a larger diversity between them which can be observed on Figure 1. The subdatasets are described in Table 1. However, the new definition of a subdataset led also to more unbalanced subdatasets: the smallest (Arvalis_8) contains only 20 images, while the biggest (ETHZ_1) contains 747 images. This provides the opportunity to possibly take advantage of the data distribution to improve model training. Each subdataset has been visually assigned to several development stage classes depending on the respective color of leaves and heads (Figure 2): postflowering, filling, filling-ripening, and ripening. Examples of the different stages are presented in Figure 2. While being approximative, this metadata is expected to improve model training.

3. Dataset Diversity Analysis

In comparison to GWHD_2020, the GWHD_2021 dataset puts emphasis on metadata documentation of the different subdatasets, as described in the discussion section of David et al. [18]. Alongside the acquisition platform, each subdataset has been reviewed and a development stage was assigned to each, except for Utokyo_3 (formerly utokyo_2) as it is a collection of images from various farmer fields and development stages. Globally, the GWHD_2021 dataset covers well all development stages ranging from postanthesis to ripening (Figure 2).

The diversity between images within the GWHD_2021 dataset was documented using the method proposed by Tolias et al. [48]. The deep learning image features were first extracted from the VGG-16 deep network pretrained on the ImageNet dataset that is considered representing well the general features of RGB images. We then selected the last layer which has a size of $14 \times 14 \times 512$ and summed it into a unique vector of 512 channels, which is then normalized. Then, the UMAP dimentionality reduction algorithm [49] was used to project representations into a 2D space. The UMAP algorithm is used to keep the existing clusters during the projection to a low-dimension space. This 2D space is expected to capture the main features of the images. Results (Figure 3) demonstrate that the test dataset used for GWHD_2020 was biased in comparison to the training dataset. The subdatasets added in 2021 populate more evenly the 2D space which is expected to improve the robustness of the models.

4. Presentation of Global Wheat Challenge 2021 (GWC 2021)

The results from the Kaggle challenge based on GWHD_2020 have been analyzed by the authors [14]. The findings emphasize that the design of a competition is critical to enable solutions that improve the robustness of the wheat head detection models. The Kaggle competition was based on a metric that was averaged across all test images, without

Table 1: The subdatasets for GWHD_2020 and GWHD_2021. The column "2020 name" indicates the name given to the subdatasets for GWHD_2020, which were split into several new subdatasets.

GWHD_2021 subdataset name	GWHD_2020 subdataset name	Owner	Country	Location	Acquisition date	Platform	Development stage	Number of images	Number of wheat head
Ethz_1	*ethz_1*	*ETHZ*	*Switzerland*	*Usask*	*06/06/2018*	*Spidercam*	*Filling*	*747*	*49603*
Rres_1	*rres_1*	*Rothamsted*	*UK*	*Rothamsted*	*13/07/2015*	*Gantry*	*Filling-ripening*	*432*	*19210*
ULiège-GxABT_1		*Uliège/ Gembloux*	*Belgium*	*Gembloux*	*28/07/2020*	*Cart*	*Ripening*	*30*	*1847*
NMBU_1		*NMBU*	*Norway*	*NMBU*	*24/07/2020*	*Cart*	*Filling*	*82*	*7345*
NMBU_2		*NMBU*	*Norway*	*NMBU*	*07/08/2020*	*Cart*	*Ripening*	*98*	*5211*
Arvalis_1	*arvalis_1*	*Arvalis*	*France*	*Gréoux*	*02/06/2018*	*Handheld*	*Postflowering*	*66*	*2935*
Arvalis_2	*arvalis_1*	*Arvalis*	*France*	*Gréoux*	*16/06/2018*	*Handheld*	*Filling*	*401*	*21003*
Arvalis_3	*arvalis_1*	*Arvalis*	*France*	*Gréoux*	*07/2018*	*Handheld*	*Filling-ripening*	*588*	*21893*
Arvalis_4	*arvalis_2*	*Arvalis*	*France*	*Gréoux*	*27/05/2019*	*Handheld*	*Filling*	*204*	*4270*
Arvalis_5	*arvalis_3*	*Arvalis*	*France*	*VLB* *	*06/06/2019*	*Handheld*	*Filling*	*448*	*8180*
Arvalis_6	*arvalis_3*	*Arvalis*	*France*	*VSC* *	*26/06/2019*	*Handheld*	*Filling-ripening*	*160*	*8698*
Arvalis_7		*Arvalis*	*France*	*VLB* *	*06/2019*	*Handheld*	*Filling-ripening*	*24*	*1247*
Arvalis_8		*Arvalis*	*France*	*VLB* *	*06/2019*	*Handheld*	*Filling-ripening*	*20*	*1062*
Arvalis_9		*Arvalis*	*France*	*VLB* *	*06/2020*	*Handheld*	*Ripening*	*32*	*1894*
Arvalis_10		*Arvalis*	*France*	*Mons*	*10/06/2020*	*Handheld*	*Filling*	*60*	*1563*
Arvalis_11		*Arvalis*	*France*	*VLB* *	*18/06/2020*	*Handheld*	*Filling*	*60*	*2818*
Arvalis_12		*Arvalis*	*France*	*Gréoux*	*15/06/2020*	*Handheld*	*Filling*	*29*	*1277*
Inrae_1	*inrae_1*	*INRAe*	*France*	*Toulouse*	*28/05/2019*	*Handheld*	*Filling-ripening*	*176*	*3634*
Usask_1	**usask_1**	**USaskatchewan**	**Canada**	**Saskatchewan**	**06/06/2018**	**Tractor**	**Filling-ripening**	**200**	**5985**
KSU_1		**Kansas State University**	**US**	**KSU**	**19/05/2016**	**Tractor**	**Postflowering**	**100**	**6435**
KSU_2		**Kansas State University**	**US**	**KSU**	**12/05/2017**	**Tractor**	**Postflowering**	**100**	**5302**
KSU_3		**Kansas State University**	**US**	**KSU**	**25/05/2017**	**Tractor**	**Filling**	**95**	**5217**
KSU_4		**Kansas State University**	**US**	**KSU**	**25/05/2017**	**Tractor**	**Ripening**	**60**	**3285**
Terraref_1		**TERRA-REF project**	**US**	**Maricopa, AZ**	**02/04/2020**	**Gantry**	**Ripening**	**144**	**3360**
Terraref_2		**TERRA-REF project**	**US**	**Maricopa, AZ**	**20/03/2020**	**Gantry**	**Filling**	**106**	**1274**
CIMMYT_1		**CIMMYT**	**Mexico**	**Ciudad Obregon**	**24/03/2020**	**Cart**	**Postflowering**	**69**	**2843**
CIMMYT_2		**CIMMYT**	**Mexico**	**Ciudad Obregon**	**19/03/2020**	**Cart**	**Postflowering**	**77**	**2771**
CIMMYT_3		**CIMMYT**	**Mexico**	**Ciudad Obregon**	**23/03/2020**	**Cart**	**Postflowering**	**60**	**1561**
Utokyo_1	utokyo_1	UTokyo	Japan	NARO-Tsukuba	22/05/2018	Cart ** –	Ripening	538	14185
Utokyo_2	utokyo_1	UTokyo	Japan	NARO-Tsukuba	22/05/2018	Cart ** –	Ripening	456	13010
Utokyo_3	utokyo_2	UTokyo	Japan	NARO-Hokkaido	Multi-years *** –	Handheld	Multiple	120	3085
Ukyoto_1		UKyoto	Japan	Kyoto	30/04/2020	Handheld	Postflowering	60	2670
NAU_1	NAU_1	NAU	China	Baima	n.a	Handheld	Postflowering	20	1240
NAU_2		NAU	China	Baima	02/05/2020	Cart	Postflowering	100	4918

Table 1: Continued.

GWHD_2021 subdataset name	GWHD_2020 subdataset name	Owner	Country	Location	Acquisition date	Platform	Development stage	Number of images	Number of wheat head
NAU_3		NAU	China	Baima	09/05/2020	Cart	Filling	100	4596
UQ_1	***uq_1***	***UQueensland***	***Australia***	***Gatton***	***12/08/2015***	***Tractor***	***Postflowering***	***22***	***640***
UQ_2	***uq_1***	***UQueensland***	***Australia***	***Gatton***	***08/09/2015***	***Tractor***	***Postflowering***	***16***	***39***
UQ_3	***uq_1***	***UQueensland***	***Australia***	***Gatton***	***15/09/2015***	***Tractor***	***Filling***	***14***	***297***
UQ_4	***uq_1***	***UQueensland***	***Australia***	***Gatton***	***01/10/2015***	***Tractor***	***Filling***	***30***	***1039***
UQ_5	***uq_1***	***UQueensland***	***Australia***	***Gatton***	***09/10/2015***	***Tractor***	***Filling-ripening***	***30***	***3680***
UQ_6	***uq_1***	***UQueensland***	***Australia***	***Gatton***	***14/10/2015***	***Tractor***	***Filling-ripening***	***30***	***1147***
UQ_7		***UQueensland***	***Australia***	***Gatton***	***06/10/2020***	***Handheld***	***Ripening***	***17***	***1335***
UQ_8		***UQueensland***	***Australia***	***McAllister***	***09/10/2020***	***Handheld***	***Ripening***	***41***	***4835***
UQ_9		***UQueensland***	***Australia***	***Brookstead***	***16/10/2020***	***Handheld***	***Filling-ripening***	***33***	***2886***
UQ_10		***UQueensland***	***Australia***	***Gatton***	***22/09/2020***	***Handheld***	***Filling-ripening***	***53***	***8629***
UQ_11		***UQueensland***	***Australia***	***Gatton***	***31/08/2020***	***Handheld***	***Postflowering***	***42***	***4345***
ARC_1		**ARC**	**Sudan**	**Wad Medani**	**03/2021**	**Handheld**	**Filling**	**30**	**888**
							Total	6515	275187

*VLB: Villiers le Bâcle; VSC: Villers-Saint-Christophe. **Utokyo_1 and Utokyo_2 were taken at the same location with different sensors. ***Utokyo_3 is a special subdataset made from images coming from a large variety of farmers in Hokaido between 2016 and 2019. Italic: Europe: bold: North America; underline: Asia; bold italic: Oceania; bold underline: Africa.

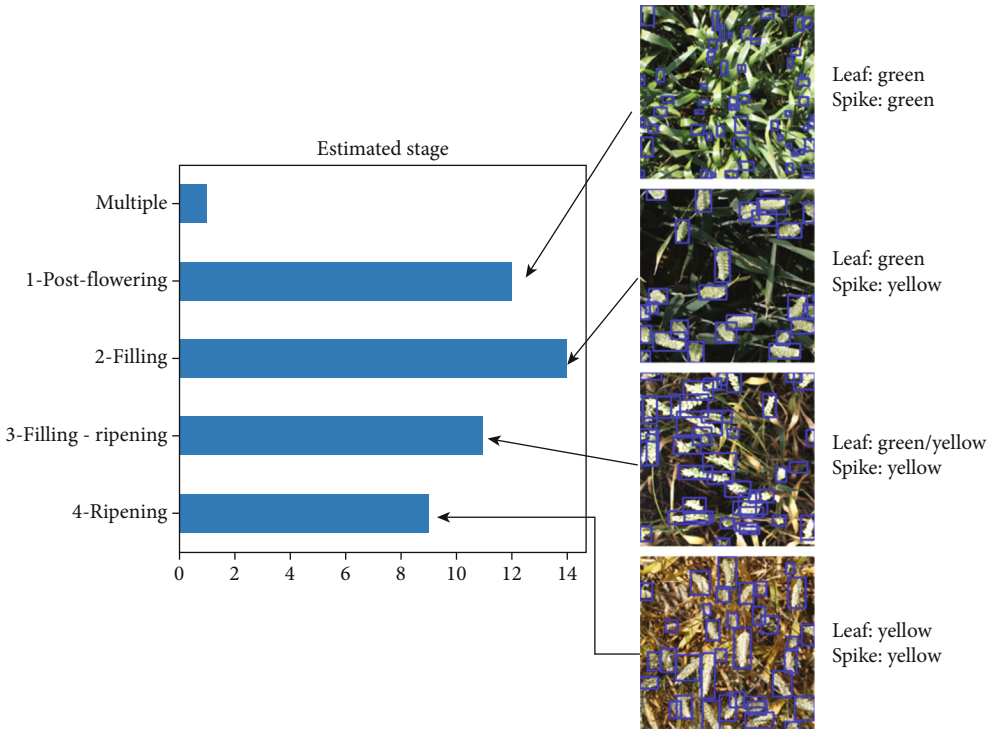


Figure 2: Distribution of the development stage. The x-axis presents the number of subdataset per development stage.

distinction for the subdatasets, and it was biased toward a strict match of the labelling. This artificially enhances the influence on the global score of the largest datasets such as utokyo_1 (now split into Utokyo_1 and Utokyo_2). Further, the metrics used to score the agreement with the labeled heads and largely used for big datasets, such as MS COCO, appear to be less efficient when some heads are labeled in a more uncertain way as it was the case in several situations depending on the development stage, illumination conditions, and head density. As a result, the weighted domain

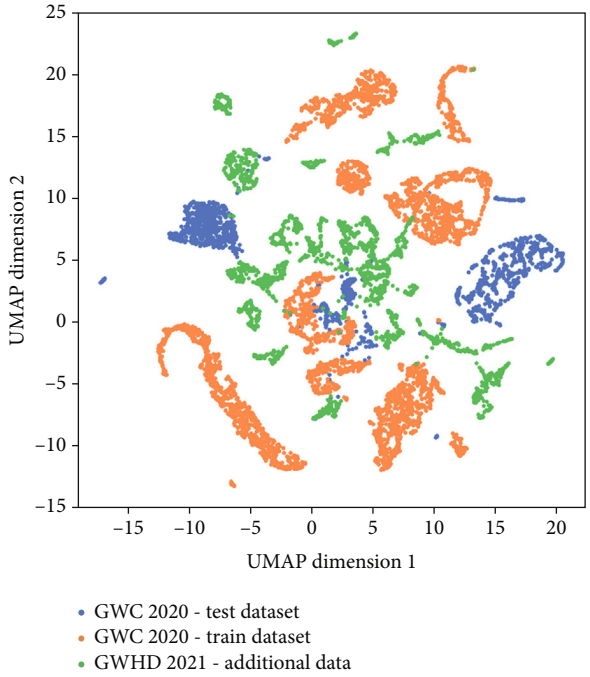


Figure 3: Distribution of the images in the two first dimensions defined by the UMAP algorithm for the GWHD 2021 dataset. The additional subdatasets as well as the training and test datasets from GWHD_2020 are represented by colors.

Table 2: Presentation of the Global Wheat Challenge 2021 results.

Solution name	WDA
randomTeamName (1st place)	0.700
David_jeon (2nd place)	0.695
SMART (2nd place)	0.695
Reference (faster-RCNN)	0.492

accuracy is proposed as a new metric [14]. The accuracy computed over image i belonging to domain d, $\mathrm{AI}_d(i)$, is classically defined as

$$\mathrm{AI}_d(i) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}, \tag{1}$$

where TP, FN, and FP are, respectively, the number of true positive, false negative, and false positive found in image i. The weighted domain accuracy (WDA) is the weighted average of all domain accuracies:

$$\mathrm{WDA} = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{n_d} * \sum_{i=1}^{n_d}\mathrm{AI}_{di}, \tag{2}$$

where D is the number of domains (subdatasets) and n_d is the number of images in domain d. The training, validation, and test datasets used are presented in Section 5.

The results of the Global Wheat Challenge 2021 are summarized in Table 2. The reference method is a faster-RCN with the same parameters than in the research paper GWHD_2020 [18] and trained on the GWHD_2021 (Global Wheat Challenge 2021 split) training dataset. The full leaderboard can be found at https://www.aicrowd.com/challenges/global-wheat-challenge-2021/leaderboards.

5. How to Use/FAQ

(i) How to download? The dataset can be download on Zenodo: https://zenodo.org/record/5092309

(ii) What is the license of the dataset? The dataset is under the MIT license, allowing for reuse without restriction

(iii) How to cite the dataset? The present paper can be cited when using the GWHD_2021 dataset. However, cite preferentially [18] for wheat head detection challenges or when discussing the difficulty to constitute a large datasets

(iv) How to benchmark? Depending on the objectives of the study, we recommend two sets of training, validation, and test (Table 3):

(a) The Global Wheat Challenge 2021 split when the dataset is used for phenotyping purpose, to allow direct comparison with the winning solutions

(b) The "GlobalWheat-WILDS" split is the one used for the WILDS paper [50]. We recommand to use the GlobalWheat-WILDS split when working on out-of-domain distribution shift problems

It is further recommended to keep the weighted domain accuracy for comparison with previous works.

Table 3: Presentation of the different splits which can be used with the GWHD_2021.

Splits	Training	Validation	Test
Global Wheat Challenge 2021	Ethz_1, Rres_1, Inrae_1, Arvalis (all), NMBU (all), ULiège-GxABT (all)	UQ_1 to UQ_6, Utokyo (all), NAU_1, Usask_1	UQ_7 to UQ_12, Ukyoto_1, NAU_2 and NAU_3, ARC_1, CIMMYT (all), KSU (all), Terraref (all)
GlobalWheat-WILDS	Ethz_1, Rres_1, Inrae_1, Arvalis (all), NMBU (all), ULiège-GxABT (all)	UQ (all), Utokyo (all), Ukyoto_1, NAU (all)	CIMMYT (all), KSU (all); Terraref (all), Usask_1, ARC_1

6. Conclusion

The second edition of the Global Wheat Head Detection, GWHD_2021, alongside the organization of a second Global Wheat Challenge is an important step for illustrating the usefulness of open and shared data across organizations to further improve high-throughput phenotyping methods. In comparison to the GWHD_2020 dataset, it represents five new countries, 22 new subdatasets, 1200 new images, and 120,000 new-labeled wheat heads. Its revised organization and additional diversity are more representative of the type of images researchers and agronomists can acquire across the world. The revised metrics used to evaluate the models during the Global Wheat Challenge 2021 can help researchers to benchmark one-class localization models on a large range of acquisition conditions. GWHD_2021 is expected to accelerate the building of robust solutions. However, progress on the representation of developing countries is still expected and we are open to new contributions from South America, Africa, and South Asia. We started to include nadir view photos from smartphones, to get a more comprehensive dataset and train reliable models for such affordable devices. Additional works are required to adapt such an approach to other vectors such as a camera mounted on unmanned aerial vehicle, or other high-resolution cameras working in other spectral domains. Further, it is planned to release wheat head masks alongside the bounding box given the very large number of boxes that already exists and provides more associated metadata.

Data Availability

The dataset is available on Zenodo (https://zenodo.org/record/5092309).

Conflicts of Interest

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

Acknowledgments

We would like to thank the company "Human in the loop", which corrected and labeled the new datasets. The help of Frederic Venault (INRAe Avignon) was also precious to check the labelled images. The work received support from ANRT for the CIFRE grant of Etienne David, cofunded by Arvalis for the project management. The labelling work was supported by several companies and projects, including Canada: The Global Institute Food Security, University of Saskatchewan which supported the organization of the competition. France: This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004 PIA #Digitag. Institut Convergences Agriculture Numérique, Hiphen supported the organization of the competition. Japan: Kubota supported the organization of the competition. Australia: Grains Research and Development Corporation (UOQ2002-008RTX machine learning applied to high-throughput feature extraction from imagery to map spatial variability and UOQ2003-011RTX INVITA—a technology and analytics platform for improving variety selection) supported competition.

References

[1] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, *Everyone wants to do the model work, not the data work: data cascades in high-stakes AI*, New York, NY, USA, 2021.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[3] T.-Y. Lin et al., "Microsoft coco: common objects in context," *European conference on computer vision*, pp. 740–755, 2014.

[4] J. A. Cruz, X. Yin, X. Liu et al., "Multi-modality imagery database for plant phenotyping," *Machine Vision and Applications*, vol. 27, no. 5, pp. 735–749, 2016.

[5] W. Guo, B. Zheng, A. B. Potgieter et al., "Aerial imagery analysis – quantifying appearance and number of sorghum heads for applications in breeding and agronomy," *Frontiers in Plant Science*, vol. 9, p. 1544, 2018.

[6] D. P. Hughes and M. Salathé, "An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing," CoRR, 2015, http://arxiv.org/abs/1511.08060.

[7] D. LeBauer et al., "Data from: TERRA-REF, an open reference data set from high resolution genomics, phenomics, and imaging sensors," *Dryad*, p. 800302508 bytes, 2020.

[8] S. Leminen Madsen, S. K. Mathiassen, M. Dyrmann, M. S. Laursen, L. C. Paz, and R. N. Jørgensen, "Open plant phenotype database of common weeds in Denmark," *Remote Sensing*, vol. 12, no. 8, p. 1246, 2020.

[9] H. Lu, Z. Cao, Y. Xiao, B. Zhuang, and C. Shen, "TasselNet: counting maize tassels in the wild via local counts regression network," *Plant Methods*, vol. 13, no. 1, p. 79, 2017.

[10] S. Madec, K. Irfan, E. David et al., *The P2S2 segmentation dataset: annotated in-field multi-crop RGB images acquired under various conditions*, Lyon, France, 2019https://hal.inrae.fr/hal-03140124.

[11] H. Scharr, M. Minervini, A. P. French et al., "Leaf segmentation in plant phenotyping: a collation study," *Machine Vision and Applications*, vol. 27, no. 4, pp. 585–606, 2016.

[12] R. Thapa, K. Zhang, N. Snavely, S. Belongie, and A. Khan, "The Plant Pathology challenge 2020 data set to classify foliar disease of apples," *Applications in Plant Sciences*, vol. 8, no. 9, article e11390, 2020.

[13] T. Wiesner-Hanks, E. L. Stewart, N. Kaczmar et al., "Image set for deep learning: field images of maize annotated with disease symptoms," *BMC Research Notes*, vol. 11, no. 1, p. 440, 2018.

[14] E. David, F. Ogidi, W. Guo, F. Baret, and I. Stavness, *Global Wheat Challenge 2020: analysis of the competition design and winning models*, 2021.

[15] N. Hani, P. Roy, and V. Isler, "MinneApple: a benchmark dataset for apple detection and segmentation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 852–858, 2020.

[16] M. Minervini, A. Fischbach, H. Scharr, and S. A. Tsaftaris, "Finely-grained annotated datasets for image-based plant phenotyping," *Pattern Recognition Letters*, vol. 81, pp. 80–89, 2016.

[17] S. A. Tsaftaris and H. Scharr, "Sharing the right data right: a symbiosis with machine learning," *Trends in Plant Science*, vol. 24, no. 2, pp. 99–102, 2019.

[18] E. David, S. Madec, P. Sadeghi-Tehran 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," *Plant Phenomics*, vol. 2020, article 3521852, 12 pages, 2020.

[19] G. Yu, Y. Wu, J. Xiao, and Y. Cao, "A novel pyramid network with feature fusion and disentanglement for object detection," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 6685954, 13 pages, 2021.

[20] T. W. Ayalew, J. R. Ubbens, and I. Stavness, "Unsupervised domain adaptation for plant organ counting," *European conference on computer vision*, pp. 330–346, 2020.

[21] M. N. Datta, Y. Rathi, and M. Eliazer, "Wheat heads detection using deep learning algorithms," *Annals of the Romanian Society for Cell Biology*, pp. 5641–5654, 2021.

[22] F. Fourati, W. S. Mseddi, and R. Attia, "Wheat head detection using deep, semi-supervised and ensemble learning," *Canadian Journal of Remote Sensing*, vol. 47, no. 2, pp. 198–208, 2021.

[23] F. Fourati, W. Souidene, and R. Attia, "An original framework for wheat head detection using deep, semi-supervised and ensemble learning within Global Wheat Head Detection (GWHD) dataset," 2020, https://arxiv.org/abs/2009.11977.

[24] A. S. Gomez, E. Aptoula, S. Parsons, and P. Bosilj, "Deep regression versus detection for counting in robotic phenotyping," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2902–2907, 2021.

[25] B. Gong, D. Ergu, Y. Cai, and B. Ma, "Real-time detection for wheat head applying deep neural network," *Sensors*, vol. 21, no. 1, p. 191, 2021.

[26] M.-X. He, P. Hao, and Y. Z. Xin, "A robust method for wheat-ear detection using UAV in natural scenes," *IEEE Access*, vol. 8, pp. 189043–189053, 2020.

[27] Y. Jiang, C. Li, R. Xu, S. Sun, J. S. Robertson, and A. H. Paterson, "DeepFlower: a deep learning-based approach to characterize flowering patterns of cotton plants in the field," *Plant Methods*, vol. 16, no. 1, p. 156, 2020.

[28] B. Jiang, J. Xia, and S. Li, "Few training data for objection detection," in *Proceedings of the 2020 4th International Conference on Electronic Information Technology and Computer Engineering*, pp. 579–584, November 2020.

[29] A. Karwande, P. Kulkarni, P. Marathe, T. Kolhe, M. Wyawahare, and P. Kulkarni, "Computer vision-based wheat grading and breed classification system: a design approach," *Machine Learning and Information Processing: Proceedings of ICMLIP 2020*, , , Springer, p. 403, 2021.

[30] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (CNN) in vegetation remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 24–49, 2021.

[31] S. Khaki, N. Safaei, H. Pham, and L. Wang, "WheatNet: a lightweight convolutional neural network for high-throughput image-based wheat head detection and counting," 2021, https://arxiv.org/abs/2103.09408.

[32] S. U. Kolhar and J. Jagtap, *Bibliometric Review on Image Based Plant Phenotyping*16.

[33] J. Li, C. Li, S. Fei et al., "Wheat ear recognition based on RetinaNet and transfer learning," *Sensors*, vol. 21, no. 14, p. 4845, 2021.

[34] L. Lucks, L. Haraké, and L. Klingbeil, *Detektion von Weizenähren mithilfe neuronaler Netze und synthetisch erzeugter Trainingsdaten*, tm-Technisches Messen, 2021.

[35] T. Misra, A. Arora, S. Marwaha et al., "Web-SpikeSegNet: deep learning framework for recognition and counting of spikes from visual images of wheat plants," *IEEE Access*, vol. 9, pp. 76235–76247, 2021.

[36] L. G. Riera, M. E. Carroll, Z. Zhang et al., "Deep multiview image fusion for soybean yield estimation in breeding applications," *Plant Phenomics*, vol. 2021, article 9846470, 12 pages, 2021.

[37] D. T. Smith, A. B. Potgieter, and S. C. Chapman, "Scaling up high-throughput phenotyping for abiotic stress selection in the field," *Theoretical and Applied Genetics*, vol. 134, no. 6, pp. 1845–1866, 2021.

[38] Y. Suzuki, D. Kuyoshi, and S. Yamane, "Transfer learning algorithm for object detection," *Bulletin of Networking, Computing, Systems, and Software*, vol. 10, no. 1, pp. 1–3, 2021.

[39] R. Trevisan, O. Pérez, N. Schmitz, B. Diers, and N. Martin, "High-throughput phenotyping of soybean maturity using time series UAV imagery and convolutional neural networks," *Remote Sensing*, vol. 12, no. 21, p. 3617, 2020.

[40] K. Velumani, R. Lopez-Lozano, S. Madec et al., "Estimates of maize plant density from UAV RGB images using Faster-RCNN detection model: impact of the spatial resolution," 2021, https://arxiv.org/abs/2105.11857.

[41] Y. Wu, Y. Hu, and L. Li, "BTWD: bag of tricks for wheat detection," in *European Conference on Computer Vision*, pp. 450–460, Springer, 2020.

[42] H. Wang, Y. Duan, Y. Shi, Y. Kato, S. Ninomiya, and W. Guo, "EasyIDP: a Python package for intermediate data processing in UAV-based plant phenotyping," *Remote Sensing*, vol. 13, no. 13, p. 2622, 2021.

[43] Y. Wang, Y. Qin, and J. Cui, "Occlusion robust wheat ear counting algorithm based on deep learning," *Frontiers in Plant Science*, vol. 12, p. 1139, 2021.

[44] B. Yang, Z. Gao, Y. Gao, and Y. Zhu, "Rapid detection and counting of wheat ears in the field using YOLOv4 with attention module," *Agronomy*, vol. 11, no. 6, p. 1202, 2021.

[45] H. Lu, L. Liu, Y. N. Li, X. M. Zhao, X. Q. Wang, and Z. G. Cao, "TasselNetV3: Explainable Plant Counting With Guided Upsampling and Background Suppression," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–15, 2021.

[46] S. Dandrifosse, A. Carlier, B. Dumont, and B. Mercatoris, "Registration and fusion of close-range multimodal wheat images in field conditions," *Remote Sensing*, vol. 13, no. 7, p. 1380, 2021.

[47] X. Wang, H. Xuan, B. Evers, S. Shrestha, R. Pless, and J. Poland, "High-throughput phenotyping with deep learning gives insight into the genetic architecture of flowering time in wheat," *GigaScience*, vol. 8, no. giz120, 2019.

[48] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," 2015, https://arxiv.org/abs/1511.05879.

[49] L. McInnes, J. Healy, and J. Melville, *UMAP: Uuniform manifold approximation and projection for dimension reduction.*, 2020.

[50] P. W. Koh, S. Sagawa, H. Marklund et al., "WILDS: a benchmark of in-the-wild distribution shifts," 2021, April 2021, https://arxiv.org/abs/2012.07421.

3.4 Conclusion

Chapter 2 demonstrated how to build process and organize a valuable dataset for crowd-sourcing wheat head detection models. The resulting dataset is one of the largest and diverse acquisition protocols, genotype, location for field phenotyping. Even for a structure such as CAPTE (INRAe / Arvalis / Hiphen), one of the main contributors of the Global Wheat Head Dataset, the gain in diversity is expected to increase the accuracy and robustness of the wheat head density models. The next step is to analyze the winning solutions provided by the two challenges, emphasising the domain shift problem.

3.5 References

[1] 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," en, *Agricultural and Forest Meteorology*, vol. 264, pp. 225–234, Jan. 15, 2019, ISSN: 0168-1923. DOI: 10.1016/j.agrformet.2018.10.013.

4 Competition design to train robust Deep Learn model: the example of the Global Wheat Challenges

4.1 Foreword

As presented in the introduction, several techniques, including data augmentation, model architecture and choice of the hyper-parameters, can affect the final robustness of a DL model. Rather than relying on a local team to experiment with each modality to get the best model, we organized open challenges to crowdsource the best solutions. Two competitions based on the Global Wheat Head Challenge 2020 and 2021 have been organized to gather the best strategies to tackle the domain shift problem. Chapter 3 explores the output of the two Global Wheat Challenges.

Challenges are conducted on platforms that link Data scientists, people with a solid Deep Learning background, with significant problems. As the competitors are expected to focus on the DL training and optimize a single metric, the design of the competition should incentivize them to solve the domain shift problem with a meticulously designed metric. Moreover, the data split should reflect the expected domain shift. In chapter 3, the performance of the winning solutions of two challenges and the impact of the competition design are analyzed.

4.2 Global Wheat Challenge 2020 and 2021: Analysis of the competition design and winning models

The following manuscript will be submitted the "GigaSciences" journal.

Global Wheat Challenge 2020 and 2021: Analysis of the competition design and winning models

Etienne David[1,2], Franklin Ogidi[3], Daniel Smith[4], Scott Chapman[4], Benoit de Solan[2], Wei Guo[5], Frederic Baret[1], Ian Stavness[3]

[1]UMR 1114 EMMAH, INRAE, Avignon, France

[2]Arvalis – Institut du Végétal, Paris, France

[3]Department of Computer Science, University of Saskatchewan, Saskatoon, Canada

[4]School of Food and Agricultural Sciences, University of Queensland, Brisbane, Australia

[5]Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan

1 Abstract

Data competitions have become a popular approach to crowdsource new data analysis methods for general and specialised data science problems. Data competitions have a rich history in plant phenotyping, and new outdoor field datasets have the potential for recent data competitions. We developed the Global Wheat Challenge as a generalisation competition in 2020 and 2021 to see if solutions for wheat head detection from field images would work in different regions around the world. In this paper, we analyse the winning challenge solutions in terms of their robustness when applied to new datasets. We found that the design of the competition influences the selection of winning solutions and provide recommendations for future competitions focusing on the more robust solutions.

2 Introduction

Crowdsourcing is an increasingly popular approach for scientists to make advances in their field by collecting diverse raw or labelled data [1]–[4], solving problems that are difficult for algorithms but easy for humans, such as protein folding [5], or accessing large-scale distributed computing power [6]. Crowdsourcing of data analysis has increased rapidly in recent years due to the popularity of Big Data challenges [7] on web platforms such as Kaggle or Codalab. In particular, problems that are amenable to machine learning approaches, such as computer vision problems, have been promoted and popularised through open competitions such as ImageNet or COCO [8], [9].

Crowdsourcing of data and analysis have also expanded to specific application areas, such as image-based plant phenotyping, where deep learning methods have been employed for plant disease classification (Albetis et al., 2017; Fuentes et al., 2017; Toda and Okura, 2019), plant and organ detection and counting (Madec et al., 2019; David et al., 2020; Ayalew, Ubbens and Stavness, no date), vegetation segmentation (Mortensen et al., 2016). In this context, crowdsourcing of data and analysis helps to connect domain experts, e.g. plant scientists, with computer and data scientists. Such collaboration with data scientists outside the traditional scope of plant phenotyping is essential to solve fundamental data science problems within the domain.

Early computer vision competitions for plant phenotyping problems [10], [11] such as leaf counting [12] and leaf segmentation [10] were highly successful. Plant phenotyping competitions have typically focused on indoor, controlled single plant images although a few recent competition have included outdoor field data such as CVPR Agriculture-Vision competition [13] or the GrassClover competition [14]. Expanding data competition to agricultural applications are crucial for facing the new challenges of global food production in the context of global change. The availability of sensors, vectors, and data for agriculture applications is rapidly expanding, while effective data interpretation pipelines are still limited. Therefore, crowdsourcing new approaches could help addressing the data interpretation challenge in plant phenotyping and agriculture.

Despite the use of deep learning in recent plant phenotyping studies, the robustness and generalizability of these methods remains an open question, particularly for small plant datasets. The issue of robustness in trained plant phenotyping models has been difficult to study due to a lack of real-world datasets (Geirhos *et al.*, 2020; David *et al.*, 2021). A large and diverse dataset is required to study the robustness problem. Most of the Deep Learning algorithms require a large set of labelled images to be trained on, such as ImageNet [8] or MS COCO [9]. The process of labelling such images is long and tedious, limiting the availability of large training datasets. The presence of diversity in the datasets is also important to study the robustness of the models. While large datasets in plant phenotyping are already available for in-door conditions [12], [17], only few comprehensive ones exist for field conditions.

Detecting wheat head under field conditions is highly desired by breeders and agronomists: it allows estimating the head density, one of the main yield components for wheat; it allows also localising plants and describing the emergence pattern with consequences on plant competition and microplot heterogeneity; it is finally the first step before further characterisation of the heads. Several studies proposed successful methods from high resolution RGB images and deep learning methods [18], [19]. However, the training and testing datasets were limited, and it is unclear if their results apply to new datasets because of the possible variation of sensors, illumination conditions, development stages and genotypes could impact the performance as described in [16].

We compiled the Global Wheat Head Dataset (GWHD) [20], [21] to study the robustness problem in plant phenotyping. The GWHD is a large and comprehensive labelled dataset for wheat head localisation. Based on the GHWD, we organised two challenges to attract a large cohort of ML practitioners to solve the wheat head detection problem: the Global Wheat Challenge 2020 (GWC_2020) on Kaggle, which took place from 4th May to 4th August 2020 and attracted up to 2245 competitors and the Global Wheat Challenge 2021 (GWC_2021) on AIcrowd, which took place from 4th May to 4th July and attracted up to 432 competitors. This paper summarises the competitions, including a description of the most successful approaches employed. We further evaluate their robustness on two additional test datasets. Finally, we discuss lessons learned from this competition and provide recommendations for future ones.

3 Material and Methods

3.1 Datasets

Three datasets are used in the study: The Global Wheat Head Dataset, the Wheat Head Frame Dataset, and the Arvalis LITERAL dataset (table 1). Only the UQ frame dataset includes both annotated images and in-field ground counting achieved over the same 0.25 x 0.25 m sampling area. The GWHD-2021 dataset have

no in-field counting, while the Arvalis-LITERAL have no image annotation. More details on the three datasets are provided below.

Table 1: Summary of the different datasets used in the study

DATASETS	USE	IMAGE ANNOTATION	IN-FIELD COUNTING	COMMENTS
GLOBAL WHEAT HEAD	Training / Validation / Test	Yes	No	Described in [21]
UQ FRAME	Test	Yes	Yes	In-field count are made on the same sampling area than the digital images
ARVALIS LITERAL	Test	No	Yes	In-field density is manually evaluated on other parts of the microplots

3.1.1 Global Wheat Head Dataset

The Global Wheat Head Detection dataset used for the competition was extensively presented in [20], [21]. It contains 6515 high-resolution RGB images representing 275.187 wheat heads from 16 institutions across 5 continents and 12 countries (Table 2). It contains 47 sub-datasets corresponding to an image acquisition session where images are acquired at a single date over a single site with a single camera system. The images were taken from one to two meters from the soil, with different high-resolution RGB cameras providing a ground sampling distance between 0.1 to 0.43 mm. The images were carefully labelled by several operators to ensure consistency and reliability. The latest version is openly available on Zenodo (Global Wheat Head Dataset 2021 | Zenodo).

Table 2: The sub-datasets for GWHD_2020 and GWHD_2021. The column "2020 name" indicates the name given to the sub-datasets for GWHD_ 2020, which were split into several new sub-datasets. Red = Europe, Blue = North America, Yellow = Asia, Green= Oceania, Orange. The table is a reproduction from the original paper.

GWHD_2021 sub-dataset name	GWHD_2020 sub-dataset name	owner	country	location	Acquisition date	Platform	Development stage	Number of images	Number of wheat head
Ethz_1	ethz_1	ETHZ	Switzerland	Usask	06/06/2018	Spidercam	Filling	747	49603
Rres_1	rres_1	Rothamsted	UK	Rothamsted	13/07/2015	Gantry	Filling - Ripening	432	19210
ULiège-GxABT_1		Uliège/Gembloux	Belgium	Gembloux	28/07/2020	Cart	Ripening	30	1847
NMBU_1		NMBU	Norway	NMBU	24/07/2020	Cart	Filling	82	7345
NMBU_2		NMBU	Norway	NMBU	07/08/2020	Cart	Ripening	98	5211
Arvalis_1	arvalis_1	Arvalis	France	Gréoux	02/06/2018	handheld	Post-flowering	66	2935
Arvalis_2	arvalis_1	Arvalis	France	Gréoux	16/06/2018	handheld	Filling	401	21003

Arvalis_3	arvalis_1	Arvalis	France	Gréoux	07/2018	handheld	Filling - Ripening	588	21893
Arvalis_4	arvalis_2	Arvalis	France	Gréoux	27/05/2019	handheld	Filling	204	4270
Arvalis_5	arvalis_3	Arvalis	France	VLB*	06/06/2019	handheld	Filling	448	8180
Arvalis_6	arvalis_3	Arvalis	France	VSC*	26/06/2019	handheld	Filling - Ripening	160	8698
Arvalis_7		Arvalis	France	VLB*	06/2019	handheld	Filling - Ripening	24	1247
Arvalis_8		Arvalis	France	VLB*	06/2019	handheld	Filling - Ripening	20	1062
Arvalis_9		Arvalis	France	VLB*	06/2020	handheld	Ripening	32	1894
Arvalis_10		Arvalis	France	Mons	10/06/2020	handheld	Filling	60	1563
Arvalis_11		Arvalis	France	VLB*	18/06/2020	handheld	Filling	60	2818
Arvalis_12		Arvalis	France	Gréoux	15/06/2020	handheld	Filling	29	1277
Inrae_1	inrae_1	INRAe	France	Toulouse	28/05/2019	handheld	Filling - Ripening	176	3634
Usask_1	usask_1	USaskatchewan	Canada	Saskatchewan	06/06/2018	Tractor	Filling - Ripening	200	5985
KSU_1		Kansas State university	US	KSU	19/05/2016	Tractor	Post-flowering	100	6435
KSU_2		Kansas State university	US	KSU	12/05/2017	Tractor	Post-flowering	100	5302
KSU_3		Kansas State university	US	KSU	25/05/2017	Tractor	Filling	95	5217
KSU_4		Kansas State university	US	KSU	25/05/2017	Tractor	Ripening	60	3285
Terraref_1		TERRA-REF project	US	Maricopa, AZ	02/04/2020	Gantry	Ripening	144	3360
Terraref_2		TERRA-REF project	US	Maricopa, AZ	20/03/2020	Gantry	Filling	106	1274
CIMMYT_1		CIMMYT	Mexico	Ciudad Obregon	24/03/2020	Cart	Post-flowering	69	2843
CIMMYT_2		CIMMYT	Mexico	Ciudad Obregon	19/03/2020	Cart	Post-flowering	77	2771
CIMMYT_3		CIMMYT	Mexico	Ciudad Obregon	23/03/2020	Cart	Post-flowering	60	1561
Utokyo_1	utokyo_1	UTokyo	Japan	NARO-Tsukuba	22/05/2018	Cart **	Ripening	538	14185
Utokyo_2	utokyo_1	UTokyo	Japan	NARO-Tsukuba	22/05/2018	Cart**	Ripening	456	13010
Utokyo_3	utokyo_2	UTokyo	Japan	NARO-Hokkaido	Multi-years***	handheld	multiple	120	3085
Ukyoto_1		UKyoto	Japan	Kyoto	30/04/2020	handheld	Post-Flowering	60	2670
NAU_1	NAU_1	NAU	China	Baima	n.a	handheld	Post-flowering	20	1240
NAU_2		NAU	China	Baima	02/05/2020	cart	Post-flowering	100	4918
NAU_3		NAU	China	Baima	09/05/2020	cart	Filling	100	4596
UQ_1	uq_1	UQueensland	Australia	Gatton	12/08/2015	Tractor	Post-flowering	22	640
UQ_2	uq_1	UQueensland	Australia	Gatton	08/09/2015	Tractor	Post-flowering	16	39
UQ_3	uq_1	UQueensland	Australia	Gatton	15/09/2015	Tractor	Filling	14	297

UQ_4	uq_1	UQueensland	Australia	Gatton	01/10/2015	Tractor	Filling	30	1039
UQ_5	uq_1	UQueensland	Australia	Gatton	09/10/2015	Tractor	Filling - Ripening	30	3680
UQ_6	uq_1	UQueensland	Australia	Gatton	14/10/2015	Tractor	Filling - Ripening	30	1147
UQ_7		UQueensland	Australia	Gatton	06/10/2020	handheld	Ripening	17	1335
UQ_8		UQueensland	Australia	McAllister	09/10/2020	handheld	Ripening	41	4835
UQ_9		UQueensland	Australia	Brookstead	16/10/2020	handheld	Filling - Ripening	33	2886
UQ_10		UQueensland	Australia	Gatton	22/09/2020	handheld	Filling - Ripening	53	8629
UQ_11		UQueensland	Australia	Gatton	31/08/2020	handheld	Post-flowering	42	4345
ARC_1		ARC	Sudan	Wad Medani	03/2021	handheld	Filling	30	888
							Total	6515	275187

3.1.2 UQ Frame Dataset

The UQ frame dataset is a collection of images acquired on three locations for a total of five dates across Australia from august to October 2020 (Table 3). For each experiment, a frame was placed in a wheat microplot and a RGB image was taken from nadir view with a Sony RX0 camera. It provides a ground sampling distance around 0.3 mm/px. Each image is rotated and cropped to the limit of the frame. Different frames were used and presented in table 2. A human operator counted the number of wheat head in-field, and each processed images were also manually labelled with bounding boxes with the same methodology as described in DAVID et al., 2020. These images belong to the UQ_7 to UQ_11 sub-datasets corresponding to single acquisition sessions used in the test dataset during the 2021 competition. A sample of the dataset is shown in figure 1.

Table 3: Description of the different sub-datasets of UQ Frame dataset

Sub-dataset name	Date of acquisition	Location	Lat (°)	Long (°)	Sowing date	Frame number	Frame size
Frame_UQ_7	31/08/2020	Gatton	-27.55	152.27	12/05/2020	42	500mm x 1000mm
Frame_UQ_8	22/09/2020	Gatton	-27.55	152.27	06/06/2020	53	500mm x 1000mm
Frame_UQ_9	06/10/2020	Gatton	-27.55	152.27	06/06/2020	17	500mm *500mm
Frame_UQ_10	09/10/2020	McAlister	-33.98	116.40	10/06/2020	40	515mm x 515mm
Frame_UQ_11	16/10/2020	Brookstead	-27.75	151.44	12/06/2020	33	515mm x 515 mm

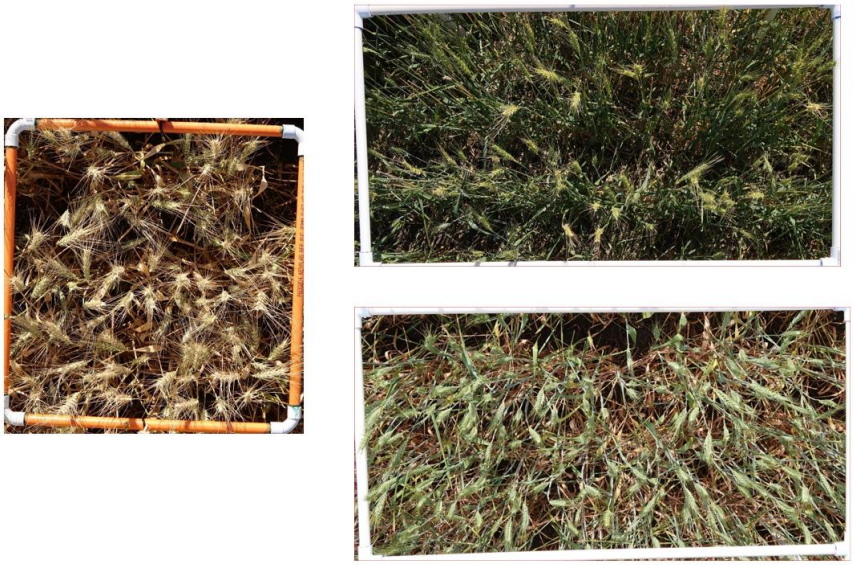


Figure 1: Sample from UQ frame dataset

3.1.3 Arvalis LITERAL Dataset

The Arvalis LITERAL Dataset is a collection of RGB images collected on five locations for a total of 6 different sessions of measurement in France from May to June 2021 (Table 4). A set of 4 images per microplot was acquired with a Sony RX0 with a resolution of 1424px x 1424 px, and an estimated canopy sampling distance (CSD) of 0.3mm. The distance between the sensor and the canopy was estimated with a stereovision algorithm and was provided with the dataset. A sample of the dataset is shown in figure 2. The head density for each image was computed as the number of heads in the image divided by the corresponding footprint area derived from the distance to the canopy and the focal length. Finally, the head density per microplot was averaged over the four images. Reference in-field head density was measured by counting the heads over two adjacent rows of 1 m length segment corresponding to a 0.35 m² sampling area.

Table 4: Presentation of the different sub-datasets of Arvalis LITERAL Dataset

Sub-dataset name	Date of acquisition	Lat(°)	Long (°)	Sowing date	Number of microplots	in field counting date and stage
Greoux	18/06/2021	43.75	5.85	28/10/2020	21	31/05/2021 (Z69)
OLM_phenohd	24/06/2021	47.90	1.52	26/10/2020	6	n.a
OLM_tpgen	17/06/2021	47.86	1.28	30/10/2020	4	23/06/2021 (Z69)
Bignan	17/06/2021	47.88	-2.75	12/11/2020	36	02/06/2021 (Z85)
Clermont	15/06/2021	45.67	2.86	30/10/2020	24	26/5/2021 (Z55)
Encrambade	04/06/2021	43.40	1.64	5/11/2020	80	01/06/2021 (Z71)

Figure 2: Sample from the Arvalis LITERAL Dataset

3.2 The Global Wheat Challenge 2020 and 2021

Despite progress in wheat head detection thanks to deep learning algorithms [18], [19], no large and diverse dataset was available to study the robustness of their solutions. The Global Wheat Head Detection 2020 [20] and 2021 [21] were the first datasets designed to match the need for more variety. The two challenges organised in 2020 and 2021 were designed to crowdsource the best solutions for wheat head detection. The experience gained with the 2020 challenge was crucial in improving both the dataset and the metrics used to compare the proposed solutions. The first challenge took place on Kaggle from 4th May to 4th August 2020 and gathered 2235 competitors with a cash prize of 15.000 USD using the GWHD_2020. The second challenge was held on AIcrowd from 4th May to the 4th of July 2021 and drew 432 contestants with a cash prize of 4.000 USD using GWHD_2021. Different aspects of a challenge organisation impact the output solutions: the cash prize, the platform and the easiness to onboard the competition greatly influence the total number of participants.

Most of the rules applied for both challenges: the competitors could access a training dataset composed of images from Europe and submit their predictions on a separate dataset consisting of images from North America, Africa, Asia and Oceania. This set was blindly split into a public test and a private test. For each submission during the competitor, the competitor could obtain a score on the public test set, but the final ranking was made once on the private test set, later called test dataset. Competitors could get their score on the public test set up to 50 times to optimize few hyperparameters, later called for this reason validation dataset. The winning solutions were expected to be opensource, with a MIT licence for the GWC_2020 and any open source licence allowing unrestricted reuse for the GWC_2021. The labels were a set of boxes around all wheat heads that can be used to train Deep Learning detection algorithms such as Faster-RCNN [22].

Two main changes occurred between the GWC_2020 and the GWC_2021 regarding the measure of robustness: the data volume and split and the metric evolved to match the original purpose of the competition.

- ***Data volume and split***. The dataset used increased from 4700 images to 6515 and from 18 to 47 sub-datasets. It represents an addition of 9 sub-datasets for training and 6 for the private test set. Further, the public and private test sets are entirely disjoint in GWC_2021, while public and private test sets were randomly drawn from the same sub-datasets. The European sub-datasets cover all necessary development stages. The diversity in terms of data owner on both test set in GWC_2021

help to avoid overfitting on specific conditions. The notion of sub-datasets, as defined in [21], was introduced to understand better the factors controlling the model robustness: a subdataset is composed of a unique acquisition session, i.e. images acquired on a single site, at a single date, with specific illumination conditions, and with the same system. However, for the Kaggle competition, the sub-datasets were not as clearly defined.

- ***The metrics used.*** The Intersection over Union (IoU) ratio is used to define the confusion matrix terms. A true positive (TP) is a labelled bounding box that matches a predicted one with an IoU ratio larger than the threshold value. A false positive (FP) is a predicted bounding box having an IoU ratio lower than the threshold with any labelled bounding box. A false negative (FN) is a labelled bounding box having an IoU ratio lower than the threshold with any predicted bounding box. In the case of GWC_2020; the accuracy at the image level, A_i, is the average of the accuracy across IoU values ranging from 0.5 to 0.75 with 0.05 steps.

$$A_i = 1/6 \sum_{IoU=0.5}^{IoU=0.75} \frac{TP(IoU)}{TP(IoU)+FP(IoU)+FN(IoU)}$$

Note that the accuracy is also the true positive rate (TPR). Then the accuracy at the image level was averaged over all the images of the test dataset to get the metric used in the GWC_2020, AA_{2020};

$$AA_{2020} = \frac{1}{n} * \sum_{i=1}^{n} A_i,$$

For GWC_2021, the accuracy is calculated for a unique IoU threshold of 0.5 to favor detection over overfitting on a particular labelling style. Further, because the test dataset was composed of several sub-datasets having very different sizes (Table 1 and Table 2), we proposed to use an average domain accuracy, AA_{2021}: during GWC_2021 competition, based on the analysis of the GWC_2020 metric (Supplementary material 1).

$$AA_{2021} = \frac{1}{D}\sum_{d=1}^{D} \frac{1}{n_d} * \sum_{i=1}^{n_d} Accuracy_{di}$$

Images from the UQ frame dataset and Arvalis LITERAL dataset are larger than the 1024px X 1024 px image size used during the challenge. Two strategies are used to handle such use cases. The baseline and GWC_2020 required a specific adaptation where the large image is split into a set of 1024px patches with a stride of 800px, representing an overlap of 22%. Models are then applied on each patch and predictions are merged with a Non-Max Suppression algorithm, using the same Intersection over Union value than used within the model (0.92 for GWC_2020 and 0.7 for Faster-RCNN). The solution GWC_2021 can automatically adapt to images with higher resolution thanks to the use of adaptive pooling layer and adaptive convolution layer.

Table 5: Summary of the different challenges

** in 2020, the public and private test set were randomly sampled from the same subdatasets*

	Platform	Number of participants	Train	Validation (Public test)	Test (Private Test)

GWC 2020	Kaggle	2235	Ethz_1 ; Rres_1; Arvalis_1 to Arvalis_6 , Inrae_1 ; Usask_1	UQ_1 to UQ_6 *; Utokyo_1 to Utokyo_3*; NAU_1*	UQ_1 to UQ_6* ; Utokyo_1 to Utokyo_3*; NAU_1*
GWC 2021	AIcrowd	427	Ethz_1; Rres_1; ULiège-GxABT_1 ; NMBU_1 ; NMBU_2 ; Arvalis_1 to Arvalis_12; Inrae_1	UQ_1 to UQ_6 ; Utokyo_1 to Utokyo_3; NAU_1; Usask_1	UQ_7 to UQ_11; ARC_1; Ukyoto_1; KSU_1 to KSU_4; Terraref_1 and Terraref_2; CIMMYT_1 to CIMMYT_3.

3.3 The models to be compared

We considered the best solutions as ranked using the average domain accuracy defined previously, and not strictly the winning models: for GWC_2020, the "Praxis" solution was selected although it was ranked second according to the GWC_2020 metrics (see Supplementary material S1); for GWC_2021 the "RandomTeamName" solution was selected and was consistently ranked first in 2021. Further, we use Faster RCNN as described in [18], [20] with an input size of 512px as the baseline solution. All solutions were trained on the GWC_2021 splits which means that the GWC_2020 was retrained from scratch from the provided code. The Faster-RCNN was trained during 30 epochs on a Nvidia geforce 3090 RTX with 24Go of RAM.

3.4 Metrics used to evaluate the winning solutions

In addition to the metrics presented above for GWC_2021, other metrics were used to compare the solutions better: the rates of false positive (FPR=FP/(TP+FP+FN)) and false negative (FNR=FN/(TP+FP+FN) to better quantify possible detection problems. Further, the Root Mean Square Error (RMSE), the relative RMSE (rRMSE), the bias andthe determination coefficient (r2) were used to quantify the head count and head density performances.

4 Results and Discussion

4.1 The GWC_2020 and GWC_2021 winning solutions

The three best solutions of the GWCs are open-source models in the form of Jupyter notebooks (Kaggle for GWC_2020) or python code (AIcrowd for GWC_2021) to reproduce the inference process. During the competition design, we expected candidates to propose domain-adaptation approaches with advanced data augmentation strategies, domain-adversarial training, along with new model architectures that could solve the difficulty of overlapping heads and small heads. Surprisingly, all winners used standard existing open-source architectures such as EfficientDet, Faster-RCNN, Yolo-v5 and Yolo-v3, without any specific domain adaptation module. The use of different architectures indicates that more than one architecture can generalise to unseen datasets. However, Yolo-v5 seems was most frequently used during GWC_2021 while it was forbidden during GWC_2020.

Part of the improved performances comes from test-time augmentation combined with the weighted boxes fusion (Solovyev, Wang and Gabruseva, 2019) that was used on the six winning solutions. Winners used several data augmentation techniques such as Mixup (Zhang *et al.*, 2018) and Mosaic augmentation, described in YoloV4 (Bochkovskiy, Wang and Liao, 2020). It is, however, difficult to assess the effect of the specific data augmentation strategies on robustness. We identified this particular topic as a question to investigate to find a better baseline strategy than the proposed Faster-RCNN. Winners all used pseudo-labelling. Pseudo-labelling (Lee, 2013) is the practice of converting predictions to labels over the test set called "pseudo-labelled" data and then fine-tuning the model with a mix of training data and pseudo-labelled data. Further, an ensemble approach was proposed by five winners over six, where several models were trained on different subsampling of the training dataset, and their solutions were fused into a single solution. Despite being widely used for data competition, these techniques are not standard in plant phenotyping.

Additional strategies were developed during the competition. First, the participants optimized few hyperparameters including the score threshold, the IoU threshold, and the image size. None of the winning solutions used the same set of hyperparameters as the proposed baseline solution. For instance, the second winner of GWC_2021 (david_jeon) upsampled the images to 1600px before prediction, which could help detect small wheat heads. Another popular approach during GWC_2020 was to generate more diversity from the training dataset based on "jigsaw puzzle" techniques. Given that some images were cropped from an original larger one, competitors have recreated images by re-cropping new patches randomly instead of applying a regular grid as in the baseline solution preprocessing step (David *et al.*, 2020).

Table 6: Summary of the winning solutions

**The score was not obtained during a competition*

	Rank	Solution name	Domain Data Augmentation	Architecture	Ensemble approach	Challenge score	Comments
Baseline		**Madec**	**No**	**Faster-RCNN**	**No**	$\boldsymbol{AA_{2021} = 0.474}$ *	
GWC_2020	1	DungNB	Mixup ; Custom mosaic	EfficientDet; FasterRCNN	Random subsampling	AA_{2020} =0.690	
	2	**OverFeat**	**Mixup, Cutmix**	**Efficentdet**	**Random subsampling**	$\boldsymbol{AA_{2020}}$ **=0.688**	Rank first with AA_{2021}
	3	Javu	Mixup	YoloV3	No	AA_{2020} =0.684	
GWC_2021	**1**	**Random TeamName**	**Mosaic**	**Yolov5**	**Domain subsampling**	$\boldsymbol{AA_{2021}}$ **=0.700**	
	2	David_jeon	Mosaic; CutMix	Yolov5	No	AA_{2021} =0.695	Model is applied on 1600 px images
	3	SMART	CutMix	Yolov4	Yes	AA_{2021} =0.695	A network is jointly trained to improve image quality

4.2 Challenges solved most of the false positives, but still miss small wheat heads

The performances measured with the average accuracy (AA_{2021}) are very heterogeneous between domains while the ranking of AA_{2021} across domains is similar for all the three approaches considered here (Figure

3). It appears that the level of complexity of an acquisition domain depends on its inner characteristics and all the solutions considered experienced difficulties in such situations. However, the GWC_2021 always beats the other solutions except for Terraref_2 and NAU_3 domains. Conversely, the baseline solution has always lower AA_{2021} values than the two winning solutions. The GWC_2020 solution show generally intermediate AA_{2021} except for few domains with values close to that of GWC_2021, and even better for Terraref 2 and NAU 3 domains.

While the results just discussed correspond to those obtained over the test dataset, it is interesting to compare them to those of the training and validation datasets: it provides some insight on the possible domain shifts and the robustness. The AA_{2021} values of the training dataset always get the best accuracy as expected. This is also observed for the false positive and false negative rates that are lower than those of the training and test datasets (Figure 3). The validation dataset used for hyperparameter optimization and pseudo-labelling has AA_{2021} values in between the training and the test datasets. The domain shift evaluated by the difference in AA_{2021} between the training and test datasets is reduced for GWC_2021, but significant for the two other solutions. The GWC_2021 solution appears more robust than the baseline and GWC_2020 models.

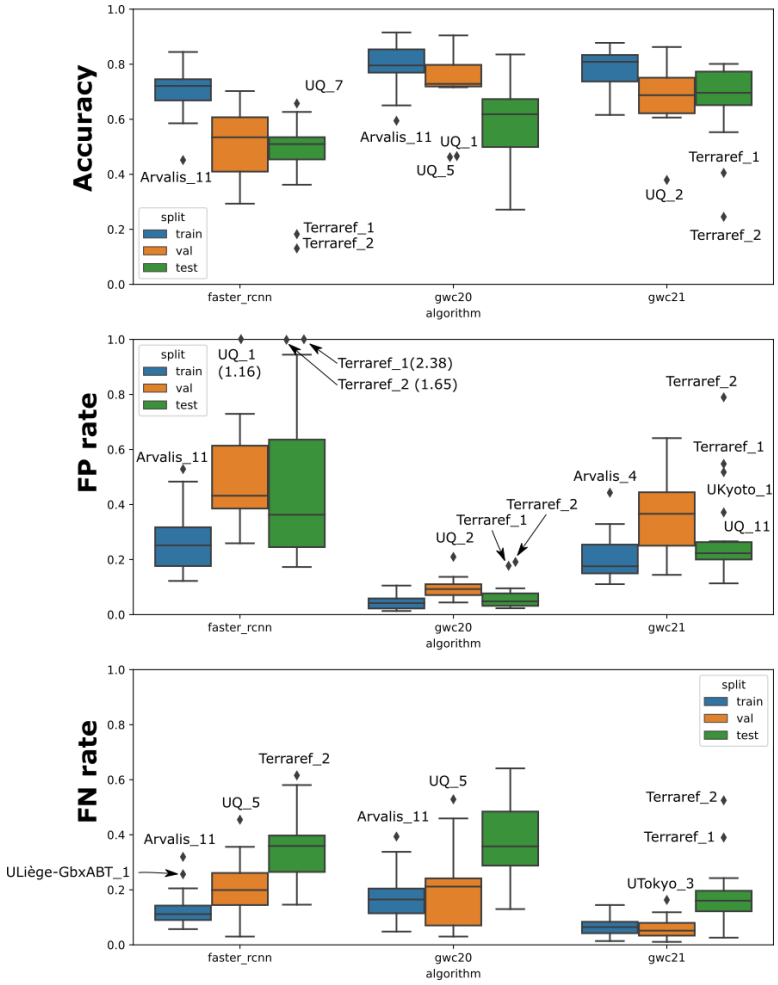


Figure 3: Detailed performance of the best solutions against the baseline. Top: Accuracy per domain; center: False positive rate (FPR); Bottom: False negative rate (FNR)

The better AA_{2021} obtained for GWC_2021 is mostly coming from a reduction of the false negatives (Figure 3 bottom): GWC_2021 gets always the lower FNR values across all the domains. The GWC_2020 and baseline solutions show similar FNRs. Conversely, the GWC_2020 gets always the smallest false positive rates, while the baseline solution shows the higher values of FPR (Figure 3, middle). The GWC_2021 solution shows intermediate, but significant values of FPRs. Combining the GWC_2021 and GWC_2020 solutions could be a possible pertinent solution when selecting the bounding boxes to be kept. This will be investigated in a future work.

Part of the low performance comes from the specificity of certain subdatasets, which are more complex than others. Some common patterns can be found: first, low resolution is a factor in the difficulty of Arvalis_4. Some subdatasets such as UQ_1 and UQ_2 contain many empty images, which lower or increase enormously the accuracy on the subdataset. The combination of wheat head bending and awns explain Arvalis_11, UQ_5 and UQ_7. Additionally to these difficulties, Terraref_1 and Terraref_2 suffer from intense illumination resulting in stark contrasts in the images. Finally, for some subdatasets (Terraref_1, Terraref_2, UQ_1, Ukyoto_1), a part of the wheat heads are not out of the stalk, despite the majority of the wheat heads being out.

The false negative rates indicate the number of heads that were not detected by the model. Even with the best model represented by GWC_2021, all the FNRs are larger than 0.10, except on four sessions. This is still a strong limitation for an accurate detection of the heads, explaining why AA_{2021} reaches values ranging from 0.2 to 0.8, with an average close to 0.7. The distribution of the size of the missing boxes (Figure 4) for all the solutions show that they generally correspond to smaller heads that are more difficult to detect. The missed heads are about 35% smaller than all the heads. If the GWC_2021 detects twice more small wheat heads, it is still missing some of the smaller heads.

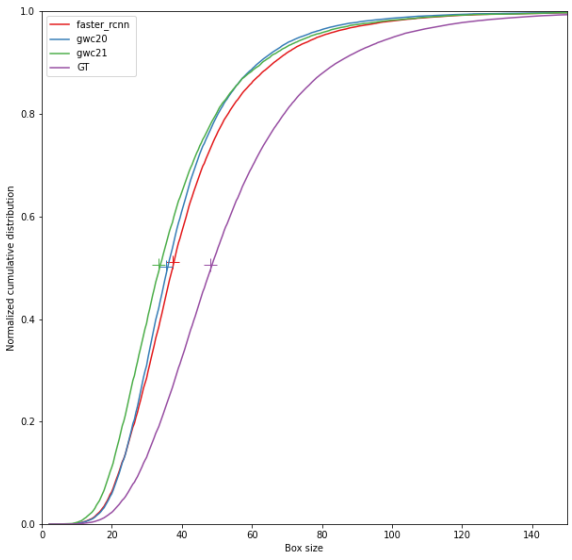


Figure 4: cumulated frequency of the size of the missing bounding boxes compared to the distribution of the size of all bounding boxes from the test set (GT).

The missing heads are often underdeveloped wheat heads or hidden wheat heads (Figure 5 top line). The difficulty of this class of small objects is the uncertainty attach to the labeling itself. Some of the labeled heads could be considered as non-wheat heads for some observer. Playing with the IoU threshold may partly contribute to improve the solution for the missed heads. However, improving the detection of the smaller wheat heads may increase the false positive rates, i.e. detecting heads that do not exist. The false positive are generally small bounding boxes that contain features close to those of actual heads (Figure5, bottom line), for instance in FP-A, a curved wheat head is detected twice. In FP-B and FP-C, a leaf part is mistaken for a wheat head. The balance between false positive and false negative is difficult to get as illustrated by the GWC_2021 that reduces the number of FNR at the expense of an increase of FPR. Conversely, GWC_2020 have low FPR at the expense of an increase of FNR. A solution that would reduce both FNR and FPR is still expected, while the balance between both terms is critical when counting heads for wheat head density estimation.

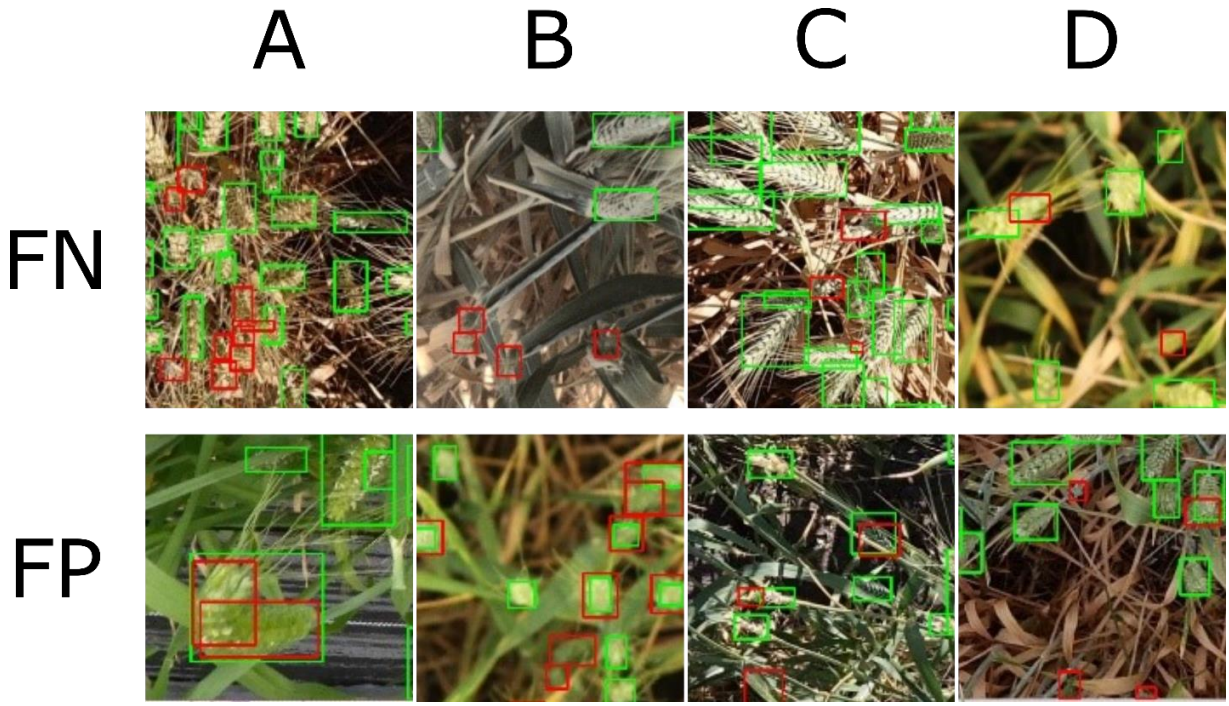


Figure 5: False negatives (FN) and positives (FP) from GWC_2021 illustrated over a random sample images of the test dataset. False negatives and positives are indicated by red bounding-boxes

GWC_2020 and GWC_2021 are very similar and use several common ideas: for the data augmentation, a mosaic augmentation, described in supplementary material S2, is applied; A pseudo-labelling procedure is applied during inference test time augmentation; Weighted boxes fusion is used. It also relies on more than one model to make the prediction. The differences are the architecture, with Yolov5 (GWC_21) instead of EfficientDet and the sampling to train the several sub-models. In GWC_2021, the selection is made at the domain level instead of the image level, explaining why the various networks used during inference have more contrasted performances. Interpreting carefully which part of the approach makes the GWC_2021 solution more robust is difficult as the hazard plays probably a significant role. However, regarding the

much lower number of competitors for GWC_2021 compared to GWC_2020, the quality of the solution is much satisfactory.

4.3 Performances for head counting

The rRMSE for each domain were calculated on all splits, including the training and validation splits (Table 7). The rRMSE varies widely across domains, from 0.66 to 2.15.

Table 7: rRMSE for the three algorithms on all domains. Colors indicate the rRMSE value: green is the lowest and red the largest

Split	Domain	rRMSE		
		Baseline	GWC_2020	GWC_2021
Training	Arvalis_1	0.23	0.153	0.215
	Arvalis_10	0.371	0.164	0.268
	Arvalis_11	0.254	0.42	0.168
	Arvalis_12	0.395	0.181	0.247
	Arvalis_2	0.139	0.144	0.152
	Arvalis_3	0.123	0.114	0.125
	Arvalis_4	0.421	0.123	0.339
	Arvalis_5	0.217	0.097	0.251
	Arvalis_6	0.084	0.155	0.129
	Arvalis_7	0.204	0.206	0.12
	Arvalis_8	0.172	0.179	0.093
	Arvalis_9	0.154	0.189	0.062
	ETHZ_1	0.073	0.164	0.134
	Inrae_1	0.145	0.083	0.185
	NMBU_1	0.079	0.313	0.097
	NMBU_2	0.241	0.224	0.124
	Rres_1	0.091	0.067	0.146
	ULiÃ¨ge-GxABT_1	0.091	0.341	0.097
Validation	NAU_1	0.239	0.291	0.472
	UQ_1	0.319	0.443	0.235
	UQ_2	2.015	0.271	1.128
	UQ_3	0.245	0.333	0.21
	UQ_4	0.264	0.275	0.122
	UQ_5	0.351	0.563	0.141
	UQ_6	0.18	0.288	0.224
	Usask_1	0.248	0.103	0.252
	Utokyo_1	0.193	0.068	0.146
	Utokyo_2	0.233	0.07	0.156
	Utokyo_3	0.511	0.246	0.308
Test	ARC_1	0.485	0.253	0.111
	CIMMYT_1	0.176	0.372	0.123
	CIMMYT_2	0.141	0.24	0.101
	CIMMYT_3	0.224	0.27	0.094
	KSU_1	0.266	0.364	0.074
	KSU_2	0.185	0.34	0.101
	KSU_3	0.177	0.323	0.173
	KSU_4	0.2	0.346	0.103
	NAU_2	0.286	0.191	0.202
	NAU_3	0.257	0.187	0.237
	Terraref_1	0.895	0.56	0.281

	Terraref_2	1.053	0.873	0.548
	UQ_10	0.166	0.462	0.157
	UQ_11	0.35	0.497	0.235
	UQ_7	0.108	0.335	0.236
	UQ_8	0.231	0.516	0.095
	UQ_9	0.212	0.532	0.102
	Ukyoto_1	0.236	0.49	0.13

A more detailed inspection of the rRMSE as a function of the model and the dataset split (Figure 6) shows as expected that the training dataset gets always the lower rRMSE values with no outliers except for GWC_2020: all the models are capable to learn the specificities of the several training domains with only small differences across the three models. The performances on the validation dataset used for hyperparameter tuning and pseudo-labeling degrade significantly, with very large outliers for the baseline and GWC_2021 models. Some domains appear difficult since they are probably too different from those of the training dataset: UQ_2 in the validation dataset shows the largest outliers for the baseline and GWC_2021 models. For the test dataset, the variability between models and domains is still large (Figure 6). The GWC_2021 model presents the lower rRMSE values and a moderate dispersion between domains. It seems the more robust solution in agreement with the detection performances presented earlier. The baseline model shows also relatively good performances except for three outliers including Terraref_2 that gets also large rRMSE for the two other models. The GWC_2020 solution presents the worst performances on the test dataset. The outliers are similar to the ones reported in 4.2.

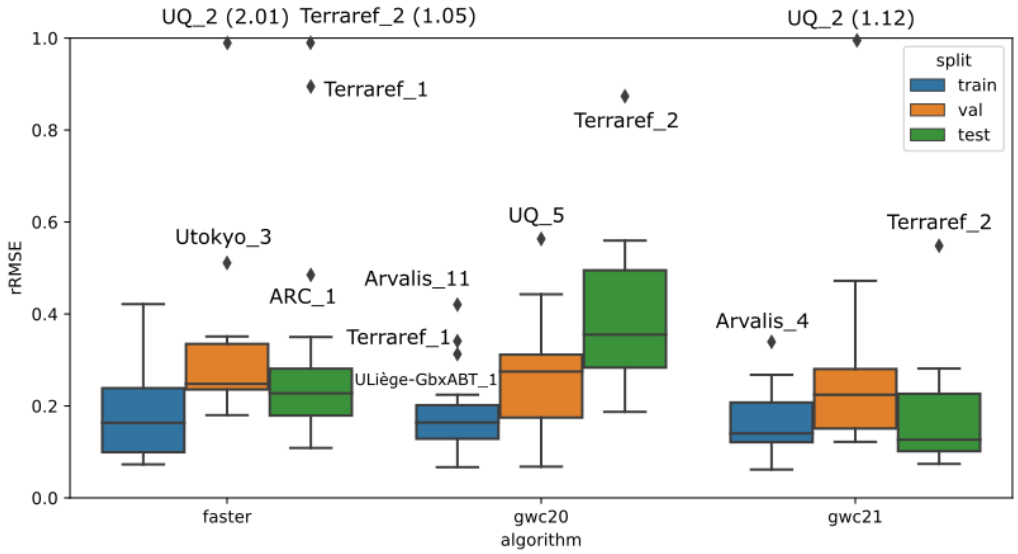


Figure 6 : Boxplot representation of the rRMSE grouped per dataset split (Train, Validation, Test) and models (Baseline, GWC2020, GWC_2021). The red line indicates the median and the blue box the 25%-75% quantiles. The whiskers extend to the most extreme (1%-99% assuming a normal distribution), while the red cross correspond to outliers. The black cross at the top indicates outliers larger than rRMSE=1, with the value on top.

The accuracy for plant detection appears not to be a good indicator of the counting performances of a model. While very poor accuracy corresponds also to high rRMSE for head counting (Figure 7), when the

accuracy is higher than 0.5, there is no relationship with the rRMSE (Figure 7). For few domains such as NAU_2 and NAU_3, the rRMSE is above 0.2 despite excellent accuracy. This is mostly due to the false negative and false positive that are strongly imbalanced (Figure 3).

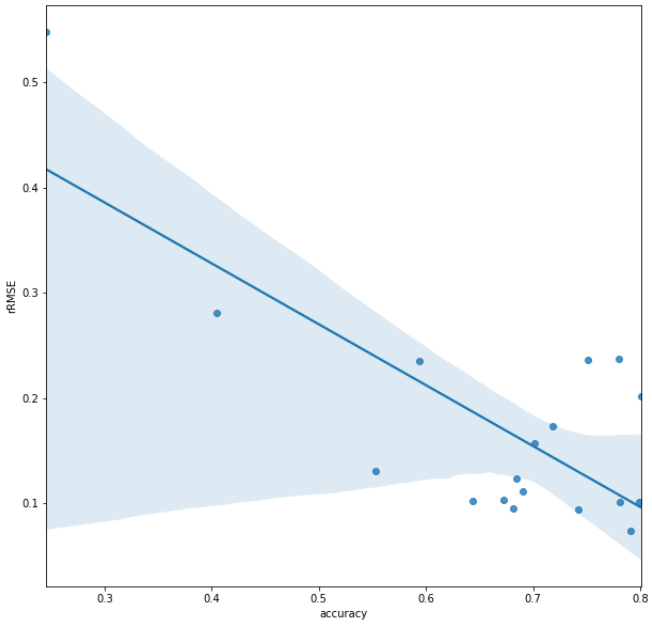


Figure 7 Relationship between accuracy and rRMSE for the test dataset and GWC_2021 model. Each of the 18 points corresponds to a specific domain.

The counting performances achieved with GWC_2021 over the test dataset are not as good as reported in other studies such as Faster-RCNN (Madec et al. [18]: rRMSE=0.06) or DeepCount (Sadeghi-Tehran et al. [23]: rRMSE=0.11). However, similar performances are observed for some domains for CIMMYT, ARC, KSU and UQ test datasets. The complexity of generalization over a wide range of conditions corresponding to the several sessions considered in the test dataset makes the detection problem much harder to solve.

4.4 Comparison with head density measurements in the field

The standard low throughput method for head density measurements is based on head counting over relatively small sampling area as described earlier in the dataset section. Two datasets, UQ frame dataset that includes images from UQ_7 to UQ_11 (Table 3) used in the validation dataset, and Literal handheld system dataset, acquired concurrently to Arvalis_1 to Arvalis_12 (used for training) but with a different RGB camera. Only the GWC_2021 model is presented here since we already demonstrated that it outperformed the baseline and GWC_2020 solutions. The performances are evaluated at the level of the microplots sampled within UQ and Literal datasets.

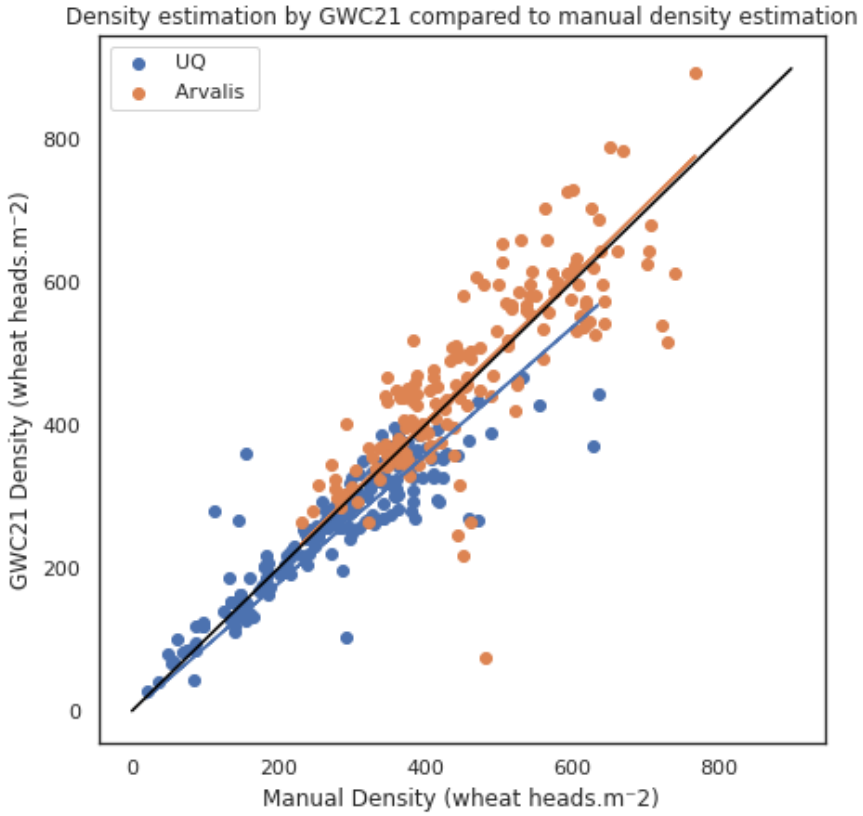


Figure 8 : Comparison between the head density measured in the field and that estimated with the three models from RGB images aggregated at the plot level. The black line is the 1:1 line. Each point corresponds to a microplot. Blue and red dts correspond respectively to UQ and Literal sessions.

Table 8 : Performances of head density estimation obtained with the GWC_2021 model.

Datasets	n	R^2	RMSE	rRMSE	Bias	slope
UQ	185	0.7323	56.58	0.2054	19.82	0.8941
Literal	163	0.633	77.14	0.1669	-9.3249	1.011
All	348	0.8135	67.00	0.1846	6.1752	0.975

There is a good agreement with the head density measured in the field at the plot level (Figure 8 and Table 8). The discrepancies seem to increase with the head density. Detecting all the heads in the image and in the field is more difficult because of possible occlusions in dense crops. The discrepancies may come from five main factors:

1. **The spatial sampling: for Literal dataset,** images and ground samples are not located at the same place, with images covering a larger area (1.4 m²) as compared to that sampled for ground-level head counting (0.7 m²).

2. **Uncertainties in the area used to compute the density.** For the UQ dataset, this uncertainty does not exist since the same frame was used to count heads in the field and to crop the images for head labeling and head prediction. For the Literal dataset, the area sampled by the image was computed from the focal length and size of the CMOS matrix that are well known and the knowledge of the distance between the top of the canopy and the camera. This distance was estimated from the distance between the camera and the canopy measured with few cm accuracy. Therefore, the error induced on the area computation should be small, on the order of few percent.
3. **Uncertainties in the model to detect heads** as seen in the previous section.
4. **Differences between heads visible in the RGB image and heads counted in the field.** The occluded heads are expected to be more frequent in dense crops (and high head density) as well as when the heads are bending as observed for many genotypes and conditions for the later maturity stages.
5. **Errors in head counting by operators in the field.** This may increase with the head density with the fatigue of operators as well as possible occlusions of heads.

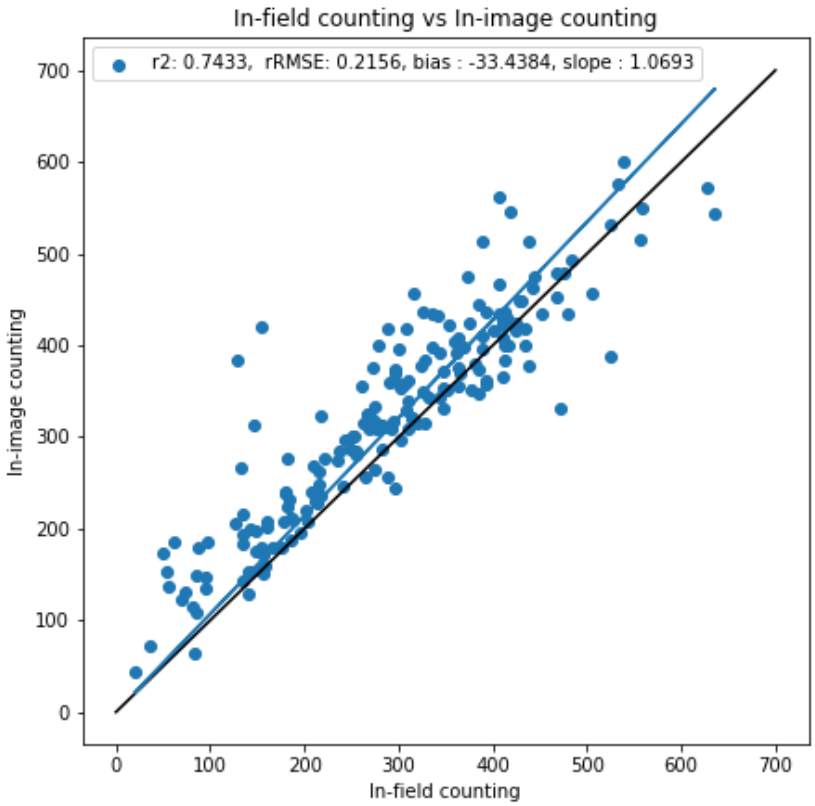


Figure 9 Comparison between the head density measured in the field and that measured on the RGB images by human operators over the same sampling area delimited by a frame. Data coming from the UQ frame dataset (n=185).

When the head density measured by an operator in the field is compared to that derived from the RGB images labelled by another operator over the same area delimited by the frame used for the UQ dataset, the

agreement is only approximative with rRMSE=21.5%. and a slight overestimation of the head density on the images (Figure 9). Although an underestimation was expected due to the occluded heads in the images as compared to what an operator can see in the field with the possibility to change the point of view and even to penetrate in the canopy, the overestimation is surprising. It may be explained by heads missed by the operator in a systematic way. When counting in the field, the operator must identify each head that are placed approximately randomly in the horizontal plane (see figures 1 and 2). It is therefore not an easy task to avoid double counting and not miss any head. Comparatively to counting on the images where the operator identifies heads incrementally with a bounding box or any other marker. This prevents double counting and limits the risk to miss heads visible on the image.

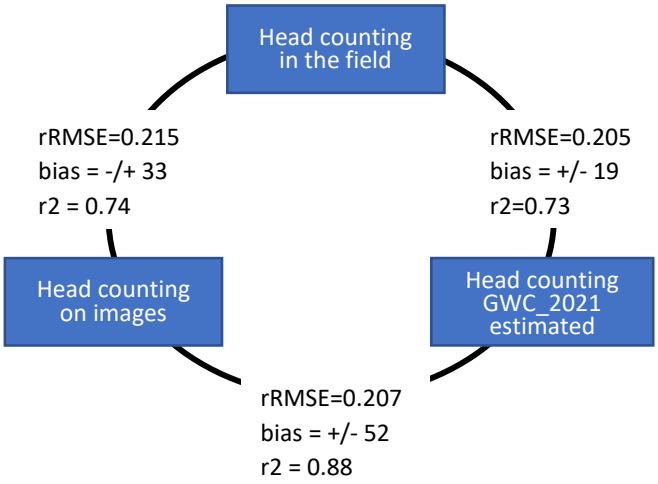


Figure 10 : Comparison of the rRMSE obtained on the UQ dataset from different ways of counting heads

A comparison between the rRMSE obtained over the UQ dataset for head counting in the frames (Figure 1) shows that the best match is observed between the head counting estimated by the GWC_2021 solution and the head counting by the operator on the images in term of correlation. Although some occluded heads may not be included in the head counting on the images, this way seems to be a better reference as compared to counting in the field where the errors made by the operator appears relatively high. Further, counting in the field is tedious and is generally done once by an operator on a limited sampling area. Conversely, counting on images can be repeated by several operators, which should improve the reliability of the result. Therefore, counting on images appears more reliable, and the confidence that we could have on counting in the field considered a reference must be altered. However, the bias between prediction and head counting in images is more significant than between the head counting in the field and the head counting on images. It means that detecting small wheat heads is a common problem for humans and the DL method. Further works are required to improve the relationship between counting in the field and counting on images by replicating on earlier stages: three dates on five were acquired during the ripping stage.

5 Conclusion

The Global Wheat Challenge 2020 and 2021 were important steps toward a robust solution to wheat head detection from high resolution RGB imagery. It complements similar initiatives focusing on plant seedling classification [11], plant pathology classification [24] , Agriculture-vision challenge [13], or plant species

recognition [25], [26]. The Global Wheat Challenge 2020 and 2021 attracted a lot of attention to a central problem in Plant Phenotyping and contributed to expose the question to a broader community, including that specialized in image processing based on artificial intelligence algorithms. It is unique in terms of the diversity of in-field situations.

The design of the competition has evolved between the two editions. Additional datasets were used in the GWC_2021, creating more diversity and more images for training. The session, i.e., images acquired with the same system at the same date and location, was recognized as a key structuring factor that defines domains. Greater attention was therefore paid on the balance between sessions in terms of the number of available images. Conversely, in the GWC_20 edition, Utokyo_1 and Utokyo_2 that contained most of the images in 2020 created an "overfit" artifact. The metrics and the split were also improved to better tackle the robustness of the models.

The three winning models of both editions were made open-source and can therefore be used by the community. The corresponding solution and weights are available on Github (https://github.com/ksnxr/GWC_solution). The proposed solutions are not very innovative in terms of model architecture, based on standard one step model such as Yolo and EfficientDet. They use specific data augmentation techniques to increase robustness, including mixup, cutmix and mosaic. Pseudo-labeling was also used to increase the diversity and size of the training dataset. Ensemble approaches were also part of the winning solutions, where several models are fused to get a more robust solution. These several elements of the winning solutions can be applied for other problems including additional traits and crops. It however requires to access to a minimal set of metadata used to define the domains [20], [21]

Solutions based on new architectures focusing on small objects that are more difficult to detect, still need to be explored. Similarly, robust algorithms such as GDRO [27] , Deep CORAL [28] should also be investigated. Further, the information of the domains/sessions was not intensively exploited during the two competitions, although it is expected to increase the robustness of the algorithms. However, the format of a challenge is perhaps not optimal for such approaches where an ensemble approach could be developed based on a series of models trained, validated, and tested on several splits using a multi-fold cross validation approach. In this case, keeping a private split for the competition will reduce the size of the training and validation datasets. Nevertheless, the Global Wheat Head Detection 2021 used for the GWC_2021 is available on the toolbox WILDS [29] to explore more advanced algorithms.

Some progress was observed on robustness in the GWC_2021 winning solution as compared to the baseline and GWC_2020 solutions both for head detection and counting. Further we demonstrated that these techniques based on high-resolution imagery were more reliable than the standard low-throughput head counting in the field by an operator: it prevents from systematic errors made by the operator and allows to increase the size of the sample used to compute the average head density at the microplot level. Further, the use of images makes the counting process more traceable as compared to counting directly in the field. Nonetheless, our study demonstrates that for well-defined conditions of acquisition, GWC_21 can be used as a replacement of manual measurement and is already used operationally both at INRAe and Arvalis. The GWC_21 solution is 14 times faster to compute compared to GWC_20 with less than one second per image on computers equipped with a Geforce 1080 GTX or a Geforce 3090 RTX Graphical processing unit board.

However, the performances on the training dataset with accuracy lower than 0.8 for detection and rRMSE larger than 0.10 for counting still needs significant improvement for the models to be used operationally

and allow deciphering small differences between genotypes or modalities. Better understanding of the impact of the acquisition system, i.e., camera type, setting, resolution is required to develop both image normalization pre-processing and efficient data augmentation techniques. Further, using bounding boxes to identify the heads may be also replaced efficiently by point identification that presents the advantage to take less time for the labeling. Further, the problems related to the score threshold and IoU are simplified in this case. Finally, at least for head counting, regression models similar to Tasselnetv2 [19] and its extension Tassenetv3 [30] or the ones used for ACID [31] may be also an interesting approach, as previous studies [32] already demonstrate superiority of regression for GWHD 2020.

6 Acknowledgements

The work received support from ANRT for the CIFRE grant of Etienne David, co-funded by Arvalis. The study was partly supported by several projects, including:

- Canada: the Global Institute Food Security, University of Saskatchewan supported the organisation of the competition
- France : PIA #Digitag Institut Convergences Agriculture Numérique , Hiphen supported the organization of the competition
- Japan: Kubota supported the organisation of the competition
- Australia: Grains Research and Development Corporation (UOQ2002-008RTX Machine learning applied to high-throughput feature extraction from imagery to map spatial variability and UOQ2003-011RTX INVITA - A technology and analytics platform for improving variety selection) supported competition

7 Bibliography

[1] H. Gao, G. Barbier, et R. Goolsby, « Harnessing the crowdsourcing power of social media for disaster relief », *IEEE Intelligent Systems*, vol. 26, nº 3, p. 10-14, 2011.

[2] R. J. Prill, J. Saez-Rodriguez, L. G. Alexopoulos, P. K. Sorger, et G. Stolovitzky, *Crowdsourcing network inference: the DREAM predictive signaling network challenge*. American Association for the Advancement of Science, 2011.

[3] A. Wiggins et K. Crowston, « From conservation to crowdsourcing: A typology of citizen science », in *2011 44th Hawaii international conference on system sciences*, 2011, p. 1-10.

[4] M. V. Giuffrida, F. Chen, H. Scharr, et S. A. Tsaftaris, « Citizen crowds and experts: observer variability in image-based plant phenotyping », *Plant Methods*, vol. 14, nº 1, p. 12, févr. 2018, doi: 10.1186/s13007-018-0278-7.

[5] B. Koepnick *et al.*, « De novo protein design by citizen scientists », *Nature*, vol. 570, nº 7761, p. 390-394, 2019.

[6] E. J. Korpela *et al.*, « Status of the UC-Berkeley SETI efforts », sept. 2011, vol. 8152. doi: 10.1117/12.894066.

[7] X. Yang, Z. Zeng, S. G. Teo, L. Wang, V. Chandrasekhar, et S. Hoi, « Deep learning for practical image recognition: Case study on kaggle competitions », in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, p. 923-931.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, et L. Fei-Fei, « Imagenet: A large-scale hierarchical image database », in *2009 IEEE conference on computer vision and pattern recognition*, 2009, p. 248-255.

[9] T.-Y. Lin *et al.*, « Microsoft coco: Common objects in context », in *European conference on computer vision*, 2014, p. 740-755.

[10] H. Scharr *et al.*, « Leaf segmentation in plant phenotyping: a collation study », *Machine vision and applications*, vol. 27, nº 4, p. 585-606, 2016.
[11] T. M. Giselsson, R. N. Jørgensen, P. K. Jensen, M. Dyrmann, et H. S. Midtiby, « A Public Image Database for Benchmark of Plant Seedling Classification Algorithms », *CoRR*, vol. abs/1711.05458, 2017, [En ligne]. Disponible sur: http://arxiv.org/abs/1711.05458
[12] M. Minervini, A. Fischbach, H. Scharr, et S. A. Tsaftaris, « Finely-grained annotated datasets for image-based plant phenotyping », *Pattern Recognition Letters*, vol. 81, p. 80-89, oct. 2016, doi: 10.1016/j.patrec.2015.10.013.
[13] M. T. Chiu *et al.*, « Agriculture-Vision: A Large Aerial Image Database for Agricultural Pattern Analysis », juin 2020.
[14] S. Skovsen *et al.*, « The GrassClover Image Dataset for Semantic and Hierarchical Species Understanding in Agriculture », in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, p. 2676-2684. doi: 10.1109/CVPRW.2019.00325.
[15] R. Geirhos *et al.*, « Shortcut Learning in Deep Neural Networks », *arXiv preprint arXiv:2004.07780*, 2020.
[16] E. David *et al.*, « Plant detection and counting from high-resolution RGB images acquired from UAVs: comparison between deep-learning and handcrafted methods with application to maize, sugar beet, and sunflower crops », *bioRxiv*, 2021, doi: 10.1101/2021.04.27.441631.
[17] S. A. Tsaftaris et H. Scharr, « Sharing the Right Data Right: A Symbiosis with Machine Learning », *Trends in Plant Science*, vol. 24, nº 2, p. 99-102, févr. 2019, doi: 10.1016/j.tplants.2018.10.016.
[18] S. Madec *et al.*, « Ear density estimation from high resolution RGB imagery using deep learning technique », *Agricultural and Forest Meteorology*, vol. 264, p. 225-234, janv. 2019, doi: 10.1016/j.agrformet.2018.10.013.
[19] H. Xiong, Z. Cao, H. Lu, S. Madec, L. Liu, et C. Shen, « TasselNetv2: in-field counting of wheat spikes with context-augmented local regression networks », *Plant Methods*, vol. 15, nº 1, p. 150, déc. 2019, doi: 10.1186/s13007-019-0537-2.
[20] E. David *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 », *Plant Phenomics*, vol. 2020, p. 3521852, août 2020, doi: 10.34133/2020/3521852.
[21] E. David *et al.*, « Global Wheat Head Dataset 2021: more diversity to improve the benchmarking of wheat head localization methods », *arXiv:2105.07660 [cs]*, juin 2021, Consulté le: sept. 05, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/2105.07660
[22] S. Ren, K. He, R. Girshick, et J. Sun, « Faster r-cnn: Towards real-time object detection with region proposal networks », in *Advances in neural information processing systems*, 2015, p. 91-99.
[23] P. Sadeghi-Tehran, N. Virlet, E. M. Ampe, P. Reyns, et M. J. Hawkesford, « DeepCount: In-field automatic quantification of wheat spikes using simple linear iterative clustering and deep convolutional neural networks », *Frontiers in Plant Science*, vol. 10, p. 1176, 2019, doi: 10.3389/fpls.2019.01176.
[24] R. Thapa, K. Zhang, N. Snavely, S. Belongie, et A. Khan, « The Plant Pathology Challenge 2020 data set to classify foliar disease of apples », *Applications in Plant Sciences*, vol. 8, nº 9, p. e11390, 2020, doi: https://doi.org/10.1002/aps3.11390.
[25] H. Goëau, P. Bonnet, et A. Joly, « Overview of lifeclef plant identification task 2019: diving into data deficient tropical countries », in *CLEF 2019-Conference and Labs of the Evaluation Forum*, 2019, vol. 2380, p. 1-13.
[26] H. Goëau, P. Bonnet, et A. Joly, « Overview of lifeclef plant identification task 2020 », 2020.
[27] S. Sagawa, P. W. Koh, T. B. Hashimoto, et P. Liang, « Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization », *arXiv:1911.08731 [cs, stat]*, avr. 2020, Consulté le: mars 12, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/1911.08731

[28] B. Sun et K. Saenko, « Deep CORAL: Correlation Alignment for Deep Domain Adaptation », *arXiv:1607.01719 [cs]*, juill. 2016, Consulté le: juill. 08, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/1607.01719

[29] P. W. Koh *et al.*, « WILDS: A Benchmark of in-the-Wild Distribution Shifts », *arXiv:2012.07421 [cs]*, mars 2021, Consulté le: avr. 14, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/2012.07421

[30] H. Lu, L. Liu, Y.-N. Li, X.-M. Zhao, X.-Q. Wang, et Z.-G. Cao, « TasselNetV3: Explainable Plant Counting With Guided Upsampling and Background Suppression », *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[31] M. P. Pound, J. A. Atkinson, D. M. Wells, T. P. Pridmore, et A. P. French, « Deep learning for multi-task plant phenotyping », in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, p. 2055-2063.

[32] A. S. Gomez, E. Aptoula, S. Parsons, et P. Bosilj, « Deep Regression Versus Detection for Counting in Robotic Phenotyping », *IEEE Robotics and Automation Letters*, vol. 6, nº 2, p. 2902-2907, 2021.

[33] H. Zhang, M. Cisse, Y. N. Dauphin, et D. Lopez-Paz, « mixup: Beyond Empirical Risk Minimization », 2018.

[34] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, et Y. Yoo, « CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features », *CoRR*, vol. abs/1905.04899, 2019, [En ligne]. Disponible sur: http://arxiv.org/abs/1905.04899

[35] M. Tan, R. Pang, et Q. V. Le, « Efficientdet: Scalable and efficient object detection », in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, p. 10781-10790.

8 Supplementary material

8.1 Limits of GWC_2020 competition design

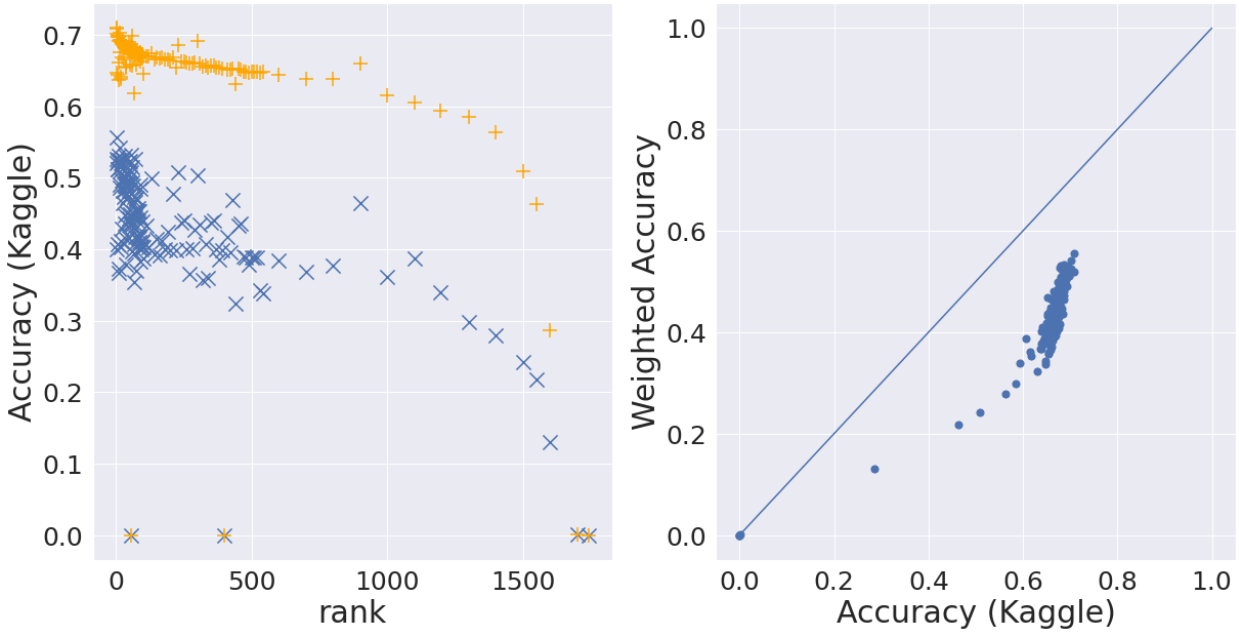


Figure 3: Presentation of the limit of the ranking used during the competition. The x-axis of Part A represents the actual ranking in the Kaggle competition while the y-axis shows the score. The scores with our re-implementation of the metric are shown in orange and the proposed simplified metric are shown in blue. Both are evaluated on the corrected private test set. Part B the score for each solution

with metric used during Kaggle against the ADA. Part C presents a simulation of the new ranking based on the new score and limited to the solutions sampled for the analysis.

The design of a competition is critical to obtain solutions that satisfy its objective. The metric proposed during the Kaggle competition presents the drawback of not having an open implementation. Our open re-implementation from scratch reaches similar but unidentical scores. These small changes can lead to severe changes in ranking. The results presented in figure 3-A reflect how the corrections and the re-implementation could have drastically changed the ranking on the Kaggle annotations.

Domain generalisation is a core problem that the dataset aims to solve but the metric chosen for the Kaggle challenge was too heavily influenced by the performance on the largest domain – utokyo_1. The use of weighted accuracy promotes solutions that have more balanced performance. Figure 3-B displays how the weighted accuracy is less saturated than the original metric with performance varying between 0.3 and 0.6 while the AA varies between 0.65 and 0.7.

Applying these results would have strongly influenced the final ranking. [Praxis](Overfeat), which has ranked 2nd and VinBigDataMedical (DungNB) which has ranked 1st would have ranked 1st and 3rd, respectively. Peculiarly, the solution that was ranked 9th would be 2nd, and the 3rd place solution would have dropped below a rank of 100. These results demonstrate the robustness of the solutions by Praxis and VinBigDataMedical despite the original metric. However, the former metric could have discouraged even more robust solutions from rising to the top. The solution by Praxis will be studied closely in the rest of the paper because it scored significantly higher (+0.03) than the second solution on the weighted accuracy. It will be called the "Kaggle solution" in the rest of the study. Henceforth, the score will be computed with the weighted accuracy on the whole public and private test set.

8.2 Impact of data augmentation: ablation study

The 20 solutions getting the highest A_{GWC_2020} were reviewed to evaluate the impact of the data augmentation strategies on the model detection performances. About 40 types of augmentation techniques were used, most of them being applied to single images, while three others require a several images to be exploited concurrently. The three multi-image augmentation techniques were selected for further analysis. About 15 single-image augmentation techniques were selected to represent the main types of techniques used, while eliminating those that were a priori not pertinent in our context such as RandomSnow or RandomSunFlare. The single-image data augmentation techniques were implemented using a popular image processing library known as Albumentations.

The multi-image augmentation techniques (Mixup, Cutmix) were used for Overfeat solution (Table 3, Figure 2). Mixup (Zhang *et al.*, 2018) and Cutmix (Yun *et al.*, 2019) both mix 2 images to form an augmented image. While Mixup uses a weighted sum of the pixels from the two images, Cutmix replaces a patch in one image with a randomly cut patch from another image. In this study the weights for the Mixup operation were set to a constant, 0.5, although it could also be drawn from a random distribution for each image. The Mosaic augmentation technique is an extension of Cutmix and takes a batch of 4 images, randomly scales each image and patches them together into a single image. Because of its close nature to

cutmix, it has not been included in the ablation study. The annotations for each image are proportionately scaled and stacked together to form the augmented annotations.

The EfficientDet-D4 [35] object detection model, pretrained on the Microsoft Common Objects in Context (MS-COCO) dataset [9] was selected as the baseline model for evaluating the effects of these data augmentation techniques on the wheat detection task using the GWHD dataset. All data augmentation strategies used in this study were applied only to the training dataset, with a probability of 0.5. Only one augmentation per group of similar operations is selected randomly for each batch.

Figure 2: Presentation of Mixup, Cutmix and Mosaic strategy

8.3 Analysis of Data Augmentation and pseudo-labelling performance

Modality	A_w	A_{GWC_2020}
Baseline	0.452787	0.654569
Baseline + Cutmix	0.476214	0.659364
Baseline + Mixup	0.367481	0.553144
Data Augmentation (DA)	0.498065	0.704271
DA + Cutmix	0.501241	0.685546
DA + Mixup	0.437714	0.686517
DA + Cutmix + Mixup	0.358226	0.508334

Table 3: Ablation study different Data Augmentation strategies

The impact of Data augmentation is important to improve the quality of the robustness. The classic data augmentation increases the weighted accuracy by 0.045, almost 10% of the baseline score. Advanced data augmentation techniques such as Cutmix [34] or Mixup [33] do not seem beneficial for the training – while Cutmix alone improves performance by 0.023, it does not add any performance compared to classic data augmentation techniques. An interesting finding is that some data augmentation can decrease the robustness – the use of Mixup always decreases the performance compared to the baseline data. Our results suggest that while the use of Cutmix or Mixup theoretically increases robustness in the use case of classification, it does not seem to translate for detection; it yields marginal gains. Conclusions are similar when using the same accuracy as in the Kaggle challenge. The result is contradictory to the popularity of such approaches during the Kaggle competition.

In our experiment, Mixup could be drawn with a probability of 0.5, while it's reduced to 0.165 in the case of the winning solution. The quality of our classic data augmentation pipeline, which is inspired by the strongest solutions could also explain the results. Our results confirm the importance of Data Augmentation for robustness but call for more careful exploration when applying usual, typically multi-image, Data Augmentation techniques. It is particularly important to remember that results on classification tasks may not translate well on detection. The use of bounding boxes may also limit the use of strategies such as Cut, Paste and Mix to increase the diversity of the data. A potential axis of research would be to generate synthetic wheat heads to increase the diversity of the data. The use of GAN or Style Transfer may be a promising solution.

4.3 Conclusion

The challenges' approach is an empirical tentative at obtaining robust models. We demonstrated the importance of the competition design on the question. The results also show the need for more in-field validation data to measure the actual error between the DL prediction and the effective number of wheat heads. Our approach on the database constitution and challenges' organization can be scaled for more traits than wheat head counting. It is, however, not tackling more conservative strategies to train robust algorithms for plant phenotyping. Data challenges' are time-consuming but rewarding in terms of optimization with a high return on value. In the chapter, we will try to tackle the problem more directly by optimizing the training for robustness.

5 GlobalWheat-Wilds: Global Wheat Head Dataset as a benchmark of in-the-wild distribution shifts

This chapter corresponds to my participation to the article "WILDS: A Benchmark of in-the-Wild Distribution Shifts", published in at ICML [1]. It focuses on the contribution of the Global Wheat Dataset to study the robustness of deep learning algorithms. Etienne David, Ian Stavness and Wei Guo are the main investigators for the contribution of Global Wheat Head Dataset 2021 (GWHD_2021) to WILDS. GWHD_2021 is the only agriculture dataset of WILDS and the only detection dataset. Therefore, it forces me to implement several changes in the WILDS codebase.

WILDS: A Benchmark of in-the-Wild Distribution Shifts

Pang Wei Koh [*1] Shiori Sagawa [*1] Henrik Marklund [1] Sang Michael Xie [1] Marvin Zhang [2]
Akshay Balsubramani [1] Weihua Hu [1] Michihiro Yasunaga [1] Richard Lanas Phillips [3] Irena Gao [1] Tony Lee [1]
Etienne David [4] Ian Stavness [5] Wei Guo [6] Berton A. Earnshaw [7] Imran S. Haque [7] Sara Beery [8]
Jure Leskovec [1] Anshul Kundaje [1] Emma Pierson [3 9] Sergey Levine [2] Chelsea Finn [1] Percy Liang [1]

Figure 22. Illustration of the WILDS paper

5.1 Introduction

5.1.1 The out-of-distribution problem

Distribution shifts—where the training distribution differs from the test distribution—can significantly degrade the accuracy of machine learning (ML) systems deployed in the wild, i.e. deployed operationally over a large diversity of situations. In this work, we consider two types of distribution shifts that are ubiquitous in real-world settings: domain generalization and subpopulation shift.

- **In domain generalization**: the training and test distributions include data from related but distinct domains. This problem arises naturally in many applications, as it is generally not possible to collect a training set that covers equally all domains of interest. For example, in medical applications, it is common to train a model on patients from few hospitals, and then deploy it more broadly to hospitals outside the training set [2]; in wildlife monitoring, an animal recognition model is trained on images from one set of camera traps to be then applied to new camera traps [3]. The wheat head detection falls into the first category of subpopulation shift.
- **In subpopulation shift**: we consider test distributions that are subpopulations of the training distribution, with the goal of keeping good performances for the worst-case subpopulation. For example, it is well-documented that standard models often perform poorly on under-represented demographics [4]; [5], and so we might seek models that can perform well on all demographic subpopulations

5.1.2 Existing ML benchmarks for domain shifts

Domain shifts have been a longstanding problem in the ML research community [6]; [7]. Earlier work studied shifts in datasets for tasks including part-of-speech tagging citemarcus93treebank, sentiment analysis [8], land cover classification [9], object recognition [10], and flow cytometry [11]. However, these datasets are not as widely used today, in part because they tend to be much smaller than modern datasets.

Despite their ubiquity in real-world deployments, these types of distribution shifts are currently underrepresented in the datasets widely used in the ML community today [12]. Recent papers have focused on object recognition datasets with shifts induced by synthetic transformations, such as ImageNet-C [13], which corrupts images with noise; the Backgrounds Challenge [14] and Waterbirds ([15], which alter image backgrounds; or Colored MNIST [16], which changes the 6 colors of MNIST digits. It is also common to use data splits or combinations of disparate datasets to induce shifts, such as generalizing to photos solely from cartoons and other stylized images in PACS [17]; generalizing to objects at different scales solely from a single scale in DeepFashion Remixed [18]; or using training and test sets with disjoint subclasses in BREEDS [19] and similar datasets [13] These existing benchmarks are useful and important testbeds for method development. As they typically target well-defined and isolated shifts, they facilitate clean analysis and controlled experimentation, e.g., studying the effect of backgrounds on image classification [14], or showing that training with added Gaussian blur improves performance on real-world blurry images [18]. Moreover, by studying how off-the-shelf models trained on standard datasets like ImageNet perform on different test datasets, we can better understand the robustness of these widely-used models [13], [18], [20]–[23]. However, as we discussed in the introduction, robustness to these synthetic shifts need not transfer to the kinds of shifts that arise in real-world deployments [22]–[24], and it is thus challenging to develop and evaluate methods for training models that are robust to real-world shifts on these datasets alone.

5.1.3 Study objectives of WILDS

In the proposed WILDS initiative, a curated benchmark of 10 datasets is compiled with evaluation metrics and train/test splits representing a broad array of distribution shifts that ML models face in the wild (Figure 23). With WILDS, we seek to complement existing benchmarks by focusing on datasets with realistic shifts across a diverse set of data modalities and applications: animal species categorization [25], tumor identification [26], bioassay prediction [27], [28], genetic perturbation classification [29], wheat head detection [30], text toxicity classification [31], land use classification [32], poverty mapping [33], sentiment analysis [34], and code completion [35], [36].

These datasets reflect natural distribution shifts arising from different cameras, hospitals, molecular scaffolds, experiments, demographics, countries, time periods, users, and codebases.

WILDS builds on extensive data-collection efforts by domain experts, who are often forced to grapple with distribution shifts to make progress in their applications. To design WILDS, we worked with the WILDS team to identify, select, and adapt datasets that fulfilled the following criteria:

- **Distribution shifts with performance drops.** The train/test splits reflect shifts that substantially degrade model performance, i.e., with a large gap between in-distribution (ID) and out-of-distribution (OOD) performances.
- **Real-world relevance.** The training/test splits and evaluation metrics are designed in conjunction with domain experts to reflect real-world scenarios. In Appendix A, we further discuss the framework we use to assess the realism of a dataset.
- **Potential leverage.** Domain shift benchmarks must be non-trivial but also possible to solve, as models cannot be expected to generalize to arbitrary distribution shifts. We constructed each WILDS dataset to have training data from multiple domains, with domain annotations and other metadata available at training time. We hope that these can be used to learn robust models: e.g., for domain generalization, one could use these annotations to learn models that are invariant to domain-specific features [37], [38], while for subpopulation shift, one could learn models that perform uniformly well across each subpopulation [15], [39].

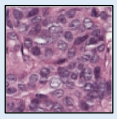
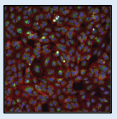
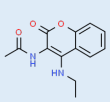
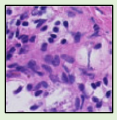
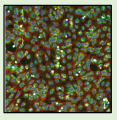
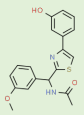
	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	**iWildCam**	**Camelyon17**	**RxRx1**	**OGB-MolPCBA**	**GlobalWheat**	**CivilComments**	**FMoW**	**PovertyMap**	**Amazon**	**Py150**
Input (x)	camera trap photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat head bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	country, rural-urban	user	git repository
# domains	323	5	51	120,084	47	16	16 x 5	23 x 2	2,586	8,421
# examples	203,029	455,954	125,510	437,929	6,515	448,000	523,846	19,669	539,502	150,000
Train example						What do Black and LGBT people have to do with bicycle licensing?			Overall a solid package that has a good quality of construction for the price.	`import numpy as np … norm=np.___`
Test example						As a Christian, I will not be patronizing any of those businesses.			I *loved* my French press, it's so perfect and came with all this fun stuff!	`import subprocess as sp p=sp.Popen() stdout=p.___`
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

Figure 23. The Wilds benchmark contains 10 datasets across a diverse set of application areas, data modalities, and dataset sizes. Each dataset comprises data from different domains, and the benchmark is set up to evaluate models on distribution shifts across these domains.

The GlobalWheat-WILDS corresponds to the domain generalization type of domain shift: we aim to generalize to test domains that are disjoint from the training domains. However, to make this problem tractable, the training and test domains are expected to be similar. A robust model should minimize the average error on the test distribution.

The aim of the study is to construct train, validation and test splits allowing to demonstrate the existence of a significant performance drops in standard models trained via empirical risk minimization (ERM) loss function. In a second step, two alternatives loss functions are tested on GlobalWheat-WILDS, to improve the robustness of the models.

5.2 Materials and methods

In this section, we detailed the contribution of the GlobalWheat-WILDS to investigate the domain shift problem.

5.2.1 The GlobalWheat-WILDS dataset

The GlobalWheat-WILDS dataset comprises 6,515 images containing 275,187 wheat heads. Most of the dataset is coming from the GWHD_2021 dataset described in chapter 2. The main difference is the repartition of the different domains in the training, validation, and test datasets. In total, 47 acquisition sessions coming from 16 research institutes across 12 countries are available. We describe the metadata and statistics of each acquisition session in Table 11. Many factors contribute to the variation in wheat appearance across acquisition sessions with substantial variation due to differences in wheat genotypes, growing conditions (e.g., planting density), illumination conditions, sensors, and measurement protocols.

We consider the domain generalization setting, where the goal is to learn models that generalize to images taken from new acquisition sessions. The task is wheat head detection, which is a single-class object detection task. Concretely, the input is an overhead outdoor image of wheat plants, with labels corresponding to bounding box coordinates that enclose the wheat heads (the spike at the top of the wheat plant containing grain), excluding the hair-like awns that may extend from the head. The domain specifies an acquisition session, which corresponds to a specific location, time, and sensor for which a set of images

were collected. Our goal is to generalize to new acquisition sessions that are unseen during training. In particular, the dataset split should capture a shift in location, with training domain made with images from countries different from those used to build the test domain.

5.2.2 The data splits used for GlobalWheat-WILDS

The goal of the GlobalWheat-WILDS, as part of the WILDS collection, is to provide a training, validation and test splits that present significant domain shifts to quantify the possible performance drop. Two main domains were defined:

- **The in domain (ID)** mages from 18 acquisition sessions in Europe (France ×13, Norway ×2, Switzerland, United Kingdom, Belgium), containing 147 957 wheat heads across 3 300 images. The ID was then split into two subdatasets by randomly drawing images from the 18 sessions:
  - ID1: used for training the models. This is the largest dataset.
  - ID2: used as a test dataset for testing the intrinsic performances in the case of Train to train setting.
- **The out of domain (OOD).** Two subdomains were identified:
  - OOD1: used for the validation, i.e. optimization of the hyperparameters. It contains Images from 7 acquisition sessions in Asia (Japan × 4, China × 3) and 1 acquisition session in Africa (Sudan), containing 44,873 wheat heads across 1,424 images.
  - OOD2: used for the test, i.e. performance evaluation. It contains Images from 11 acquisition sessions in Australia and 10 acquisition sessions in America (USA × 6, Mexico × 3, Canada), containing 66,905 wheat heads across 1,434 images.

The several splits were differently used to serve the two main objectives targeted:

- **Evaluating the performance drop.** In this case, two models were compared:
  - **Train to test**: the model is trained over ID1, validated on OOD1, and tested on OOD2. The separation between OD1 and OD2 was preferred to provide more robustness of the model by allowing optimizing the hyperparameters on an OOD dataset not used for testing the performances.
  - **Mixed to test**: the model is trained over a mix of ID1 and OOD2, validated on OOD1, and tested on OOD2. The hyperparameters used as the same than train to test.
- **Evaluating algorithms** expected to improve robustness. The model is trained on ID1 with validation on OOD1 and tested on ID2 (Train to train) and OOD2 (Train to test).

The following table 2 describes the details of the splits.

5.2.3 The detection model and hyperparameters

For the GlobalWheat-WILDS, we use the Faster-RCNN detection model [40], which has been successfully applied to the wheat head localization problem [30], [41] To train, we fine-tune a model pre-trained with ImageNet, using a batch size of 4, a learning rate of 10^{-5} , and weight decay of 10^{-3} for 10 epochs with early stopping. The hyperparameters were chosen from a grid search over learning rates 10^{-6} , 10^{-5} , 10^{-4} and weight decays 0, 10^{-4} , 10^{-3}. We report results aggregated over 3 random seeds.

5.2.4 Training algorithms used

Additionally to the study of the performance drop, we compared several training algorithms used during the learning process, some being designed to get a more robust model.

- **Empirical Risk Minimization** (Equation 1) is the classic training algorithm. The loss calculated to train a model is minimized equally for all examples (i.e. images in the case of GlobalWheat-wilds) of one batch. $h(x_i)$ represents the prediction of the DL

		Domain	ID		OOD	
		Split #	ID1	ID2	OOD1	OOD2
		# sessions	18		8	21
		# images	2943	357	1424	1434
		# labels	131864	16093	44873	66905
Performance drop	Train to Test	Train	100%			
		Validation			100%	
		Test				50%
	Mixed to Test	Train	76%			50%
		Validation			*	
		Test				50%
Algorithm evaluation		Train	100%			
		Validation			100%	
		Test		100%		100%

Table 2. The splits used for the performance drop and robustness experiments. (*) the hyperparameters used for the Mixed to Test experiment were the same as those optimized for the Train to Test experiment.

algorithm h on x_i, y_i is the corresponding label and L is the loss function.

$$R_{emp}(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(x_i),y_i) \quad (1)$$

- **Group Distributional Robustness Optimization (Group DRO)** algorithm (Equation 2) is proposed by Sagawa [42] as an alternative to ERM to minimize the loss on the worst group for each batch in contrast to the ERM algorithm. It is equivalent to calculate the empirical risk for each group (g_1,...,g_n) and minimize the maximum.

$$R_{GDRO}(h) = \max(R_{emp}(h_{g_1}), ..., R_{emp}(h_{g_n})) \quad (2)$$

- **Deep CORAL** [43] proposes to train Deep Learning model by adding a second loss term which forces to harmonize the features statistics of different domains. In practice, it minimizes the difference between the covariance matrices of two different groups in each batch. It has not been used in a detection use case to our knowledge. To adapt the algorithm, we propose to use the features of the encoder part of Faster-RCNN, used by the RPN, to calculate the Deep CORAL objective. Our framework is summarized on figure 24. Note that the encoder used for Deep CORAL was applied here on dimensions superior to 10.000, while Deep CORAL was designed for classification use case with a limited dimensionality, between 512 and 4096.

5.2.5 Metrics used to evaluate the performances

We used the average accuracy similarly to what was proposed previously for the Global Wheat Challenges with the dataset [44] . The accuracy of a bounding box detection is measured at a fixed Intersection over Union (IoU) threshold of 0.5. This permissive IoU threshold of 0.5 was selected because of the uncertainties regarding the precise outline of wheat head instances due to the stem and awns extending from the head. The accuracy of an image is computed as $\frac{TP}{TP+FN+FP}$, where TP is the number of true positives, which are ground-truth bounding boxes that have and $IoU > 0.5$ with some predicted bounding box; FN is the number of false negatives, which are ground-truth bounding boxes that have an $IoU < 0.5$ with any predicted bounding box; FP is the number of false positives, which are predicted bounding boxes that have an $IoU < 0.5$ with any ground-truth bounding box. We

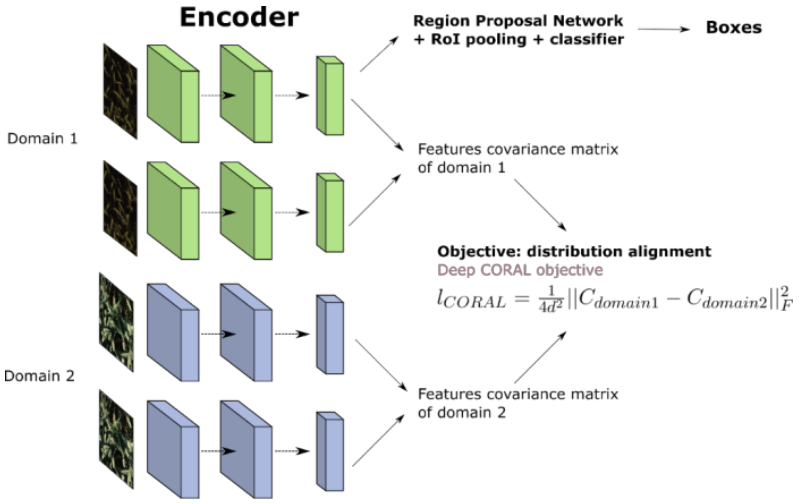


Figure 24. Deep CORAL loss objective

first computed the accuracy within each image. We then computed the average accuracy for each acquisition session by averaging the per-image accuracy, and finally averaged the accuracies of each acquisition session. Such average accuracy presents the advantage to get a more even contribution from the several sessions that show a large variability in their number of images (from 17 to 200 images in the test set). The average accuracy was preferred to the worst-case accuracy because few acquisition sessions were more difficult and yielded a poor representation of the actual model performances.

5.3 Results and Discussion

5.3.1 The performance drop is large in plant phenotyping

The per domain average accuracy of the models is very variable, with larger dispersion for the Train to test setting (Figure 25, left). The average accuracy of the mixed-to-test setting is significantly higher (Ac=63.3) than that obtained with the train to test setting (Ac=49.6). The mixed-to-test setting allows the model to learn both on ID and OOD, even if the OOD images of the train dataset represent only 24% of the total training images (but 50% of the OOD2 dataset, the other 50% being used for test, Table 2). The hyperparameters were optimized on the OOD1 dataset for the Train to test setting. The same hyperparameters were used for the Mixed to train setting. Consequently, the performances of the Mixed to train setting could have been improved if the hyperparameters were retuned on OOD1.

The performances for the two settings are very correlated (Figure 25, right). There are easy domains where the accuracy is high for both settings, with only small performance drop. Conversely, there are difficult domains for the two settings such as Terraref, because of the strong contrast in its images and the presence of partly occluded heads. The KSU sessions have also relatively poor accuracy with a large performance drop, may be due to differences in the development stage when the images were acquired. However, it is difficult to assign the performance drop to a particular factor including genotype, sensor and acquisition protocol used. Some artifacts may also contribute to make the performance drop more variable: some domains in OOD2 have a relatively small number of images which may contribute to some unbalanced distribution of the image features between the 50% OOD2 used for training the Mixed to test setting and the 50% other images used though some of the domains to compute the average accuracy. Further, a high variance across images is observed for UQ_2 and UQ_3 belonging to OOD2. It is due to the high proportion of empty images (88% for UQ_2 and 57% for UQ_3) that are accuracy scored either 0% or

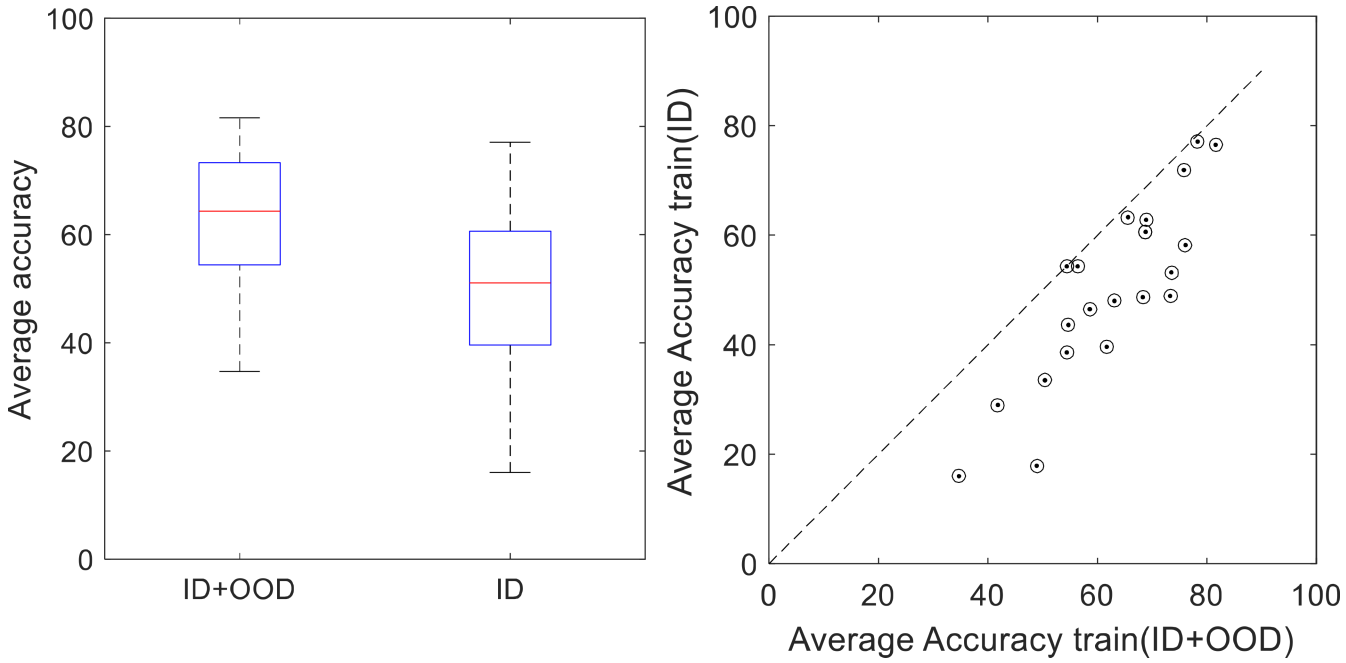


Figure 25. On the left, distribution of the average accuracy evaluated on the test (OOD) dataset when the model is trained over the mixed to train (ID+OOD) and train to test (ID) dataset. Box plot representation with the median indicated by the red line. The box contains the percentiles between 25% and 75%, while the whiskers extend to the 1%-99% percentiles. On the right, relationships between the average accuracy for each domain of the model trained on ID+OOD dataset and that trained on the ID dataset. The dashed line is the 1:1 line.

Loss function	ID2	OOD1	OOD2
ERM	77.1	68.6	51.2
Group DRO	76.2	66.2	47.9
CORAL	75.4	64.9	47.2

Table 3. Average accuracy computed over several domains (ID1, OOD1, OOD2) for the model trained on ID1 and validated on OOD1 with different loss functions (ERM, Group DRO, CORAL).

100% depending on the presence or not of false positives and negatives.

5.3.2 Alternative training algorithms do not improve robustness

The several training algorithms and loss functions tested provide relatively similar average accuracy over the three test domains considered. Further, the ERM shows always slightly higher performances as compared to Group DRO and CORAL. As expected, the average accuracy with ERM loss function evaluated on 100% of OOD2 (Ac=51.2) is consistent with the value reported for the same model over 50% of OOD2 (Ac=49.6, Figure 25). The performance drop is significant when comparing the average accuracy obtained over ID2, and in a lesser way on OOD1. However, both the group DRO and CORAL models do not reduce this performance drop despite their promises claimed by their authors. This results obtained over GlobalWheat-WILDS is consistent with those reported for most of the other WILDS datasets as demonstrated in table 3. Additionally, to the empirical results, theorical arguments are made against the IRM loss function not adapted for GlobalWheat-WILDS, but used for the other WILDS [1]. We hope that the WILDS benchmark can provide a clear framework to measure gain of robust training algorithm.

5.4 Conclusion

The inclusion of the wheat head detection problem within the WILDS initiative was an opportunity to investigate the robustness problem of deep learning approaches within a wider scientific community and to benefit from their theoretical advances. Our results demonstrate the existence of the domain shift with performance drop, which reinforces our previous conclusions on plant detection and counting from UAV observations [45]. More work is however needed to better understand the possible causes of the performance drop, either by the image features extracted by deep learning approaches or using the known metainformation that document factors such as the illumination conditions, image quality, development stage, genotype features. While this issue is not well covered in the other detection problems considered in WILDS, it appears to be also the case for the classification and segmentation problems investigated in WILDS. The domain shift and corresponding performance drop is therefore expected to impact any traits derived from a machine learning approach.

Our results also demonstrate the lack of effectiveness of some training algorithms claimed to improve the robustness of the model. The robustness seems to be better improved using ensemble approaches where several models trained differently are combined to provide a consensus solution as demonstrated in the Global Wheat Challenge [44]. Some questions are also still open: how to optimize the selection of the training, validation and test datasets? How to define in-domain (ID) and out of domain (OOD) datasets?

The use of Generative Adversarial Networks [46]–[48] could be also a solution to transfer the conditions of a domain to standard ones. Data pre-processing could also contribute to this standardization, particularly regarding the illumination conditions and color distributions. Finally, data augmentation that was not considered in WILDS may contribute to create some additional diversity and improve the robustness of the models as demonstrated in the Global Wheat Challenges. However, the growth in size and diversity of the available datasets with images labeled or not appears to be still a safe way for solving collectively a given problem and get robust models that can scale operationally over most of the situations encountered.

5.5 References

[1] P. W. Koh, S. Sagawa, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *International Conference on Machine Learning*, PMLR, 2021, pp. 5637–5664.

[2] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," in *PLOS Medicine*, 2018.

[3] S. Beery, G. V. Horn, and P. Perona, "Recognition in terra incognita," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 456–473.

[4] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 77–91.

[5] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Science*, vol. 117, no. 14, pp. 7684–7689, 2020.

[6] D. J. Hand, "Classifier technology and the illusion of progress," *Statistical science*, pp. 1–14, 2006.

[7] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. The MIT Press, 2009.

[8] J. Blitzer and F. Pereira, "Domain adaptation of natural language processing systems," *University of Pennsylvania*, 2007.

[9] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 770–787, 2009.

[10] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European Conference on Computer Vision*, 2010, pp. 213–226.

[11] G. Blanchard, G. Lee, and C. Scott, "Generalizing from several related classification tasks to a new unlabeled sample," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2011, pp. 2178–2186.

[12] R. Geirhos, J. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *arXiv preprint arXiv:2004.07780*, 2020.

[13] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations (ICLR)*, 2019.

[14] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry, "Noise or signal: The role of image backgrounds in object recognition," *arXiv preprint arXiv:2006.09994*, 2020.

[15] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in *International Conference on Learning Representations (ICLR)*, 2020.

[16] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[17] D. Li, Y. Yang, Y. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5542–5550.

[18] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: A critical analysis of out-of-distribution generalization," *arXiv preprint arXiv:2006.16241*, 2020.

[19] S. Santurkar, D. Tsipras, and A. Madry, "Breeds: Benchmarks for subpopulation shift," *arXiv*, 2020.

[20] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," *Advances in Neural Information Processing Systems*, vol. 31, pp. 7538–7550, 2018.

[21] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet classifiers generalize to ImageNet?" In *International Conference on Machine Learning (ICML)*, 2019.

[22] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, "Measuring robustness to natural distribution shifts in image classification," *arXiv preprint arXiv:2007.00644*, 2020.

[23] J. Djolonga, J. Yung, M. Tschannen, R. Romijnders, L. Beyer, A. Kolesnikov, J. Puigcerver, M. Minderer, A. D'Amour, D. Moldovan, *et al.*, "On robustness and transferability of convolutional neural networks," *arXiv preprint arXiv:2007.08558*, 2020.

[24] S. M. Xie, A. Kumar, R. Jones, F. Khani, T. Ma, and P. Liang, "In-N-Out: Pre-training and self-training using auxiliary information for out-of-distribution robustness," *arXiv*, 2020.

[25] S. Beery, E. Cole, and A. Gjoka, "The iWildCam 2020 competition dataset," *arXiv preprint arXiv:2004.10340*, 2020.

[26] P. Bandi, O. Geessink, Q. Manson, M. V. Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong, *et al.*, "From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 550–560, 2018.

[27] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "Moleculenet: A benchmark for molecular machine learning," *Chemical Science*, vol. 9, no. 2, pp. 513–530, 2018.

[28] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open graph benchmark: Datasets for machine learning on graphs," *arXiv preprint arXiv:2005.00687*, 2020.

[29] J. Taylor, B. Earnshaw, B. Mabey, M. Victors, and J. Yosinski, "Rxrx1: An image set for cellular morphological variation across many experimental batches.," in *International Conference on Learning Representations (ICLR), AI for Social Good Workshop*, 2019.

[30] E. David, S. Madec, P. Sadeghi-Tehran, H. Aasen, B. Zheng, S. Liu, N. Kirchgessner, G. Ishikawa, K. Nagasawa, M. A. Badhon, C. Pozniak, B. de Solan, A. Hund, S. C. Chapman, F. Baret, I. Stavness, and W. 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*, vol. 2020, 2020.

[31] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman, "Nuanced metrics for measuring unintended bias with real data for text classification," in *WWW*, 2019, pp. 491–500.

[32] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[33] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke, "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa," *Nature Communications*, vol. 11, 2020.

[34] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 188–197.

[35] V. Raychev, P. Bielik, and M. Vechev, "Probabilistic model for code with decision trees," *ACM SIGPLAN Notices*, 2016.

[36] S. Lu, D. Guo, S. Ren, J. Huang, A. Svyatkovskiy, A. Blanco, C. Clement, D. Drain, D. Jiang, D. Tang, G. Li, L. Zhou, L. Shou, L. Zhou, M. Tufano, M. Gong, M. Zhou, N. Duan, N. Sundaresan, S. K. Deng, S. Fu, and S. Liu, "Codexglue: A machine learning benchmark dataset for code understanding and generation," *arXiv preprint arXiv:2102.04664*, 2021.

[37] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *European conference on computer vision*, 2016, pp. 443–450.

[38] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research (JMLR)*, vol. 17, 2016.

[39] W. Hu, G. Niu, I. Sato, and M. Sugiyama, "Does distributionally robust supervised learning give robust classifiers?" In *International Conference on Machine Learning (ICML)*, 2018.

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2015, pp. 91–99.

[41] 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," en, *Agricultural and Forest Meteorology*, vol. 264, pp. 225–234, Jan. 15, 2019, ISSN: 0168-1923. DOI: 10.1016/j.agrformet.2018.10.013.

[42] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," en, *arXiv:1911.08731 [cs, stat]*, Apr. 2, 2020. [Online]. Available: http://arxiv.org/abs/1911.08731.

[43] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," *arXiv:1607.01719 [cs]*, Jul. 6, 2016. [Online]. Available: http://arxiv.org/abs/1607.01719.

[44] E. David, M. Serouart, D. Smith, S. Madec, K. Velumani, S. Liu, X. Wang, F. P. Espinosa, S. Shafiee, I. S. A. Tahir, H. Tsujimoto, S. Nasuda, B. Zheng, N. Kichgessner, H. Aasen, A. Hund, P. Sadhegi-Tehran, K. Nagasawa, G. Ishikawa, S. Dandrifosse, A. Carlier, B. Mercatoris, K. Kuroki, H. Wang, M. Ishii, M. A. Badhon, C. Pozniak, D. S. LeBauer, M. Lilimo, J. Poland, S. Chapman, B. de Solan, F. Baret, I. Stavness, and W. Guo, *Global wheat head dataset 2021: An update to improve the benchmarking wheat head localization with more diversity*, 2021. arXiv: 2105.07660 [cs.CV].

[45] E. David, G. Daubige, F. Joudelat, P. Burger, A. Comar, B. De solan, and F. Baret, "Plant detection and counting from high-resolution rgb images acquired from uavs: Comparison between deep-learning and handcrafted methods with application to maize, sugar beet, and sunflower crops," *bioRxiv*, 2021. DOI: 10.1101/2021.04.27.441631. [Online]. Available: https://www.biorxiv.org/content/early/2021/04/28/2021.04.27.441631.

[46] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial networks*, 2014. arXiv: 1406.2661 [stat.ML].

[47] S.-W. Huang, C.-T. Lin, S.-P. Chen, Y.-Y. Wu, P.-H. Hsu, and S.-H. Lai, "Auggan: Cross domain adaptation with gan-based data augmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep. 2018.

[48] W. Zhang, K. Chen, J. Wang, Y. Shi, and W. Guo, "Easy domain adaptation method for filling the species gap in deep learning-based fruit detection," en, *Horticulture Research*, vol. 8, no. 1, pp. 1–13, Jun. 1, 2021, ISSN: 2052-7276. DOI: 10.1038/s41438-021-00553-8.

6 Conclusion and perspectives

This work focuses on the interpretation of phenotyping measurements to extract pertinent traits for breeders, ecophysiologists, and agronomists. While data acquisition currently reached a high level of maturity, the estimation of traits from the row measurements is still recognized as a bottleneck before within the high-throughput phenotyping systems. The recent emergence of deep learning (DL) methods was a breakthrough in data interpretation. It allows not only to improve the traitestimation performances, but also to extract traits which were impossible to access with standard computer vision and machine learning techniques . Our study focused on plants and organ identification and counting that are the main components of the yield and therefore highly desired within the set of high-throughput phenotyping traits. Identifying plants is not only a way to count them but also to localize them and describe their local environment while allowing further characterization of these individual objects. For plant counting, we demonstrated that DL techniques generally outperform more standard computer vision techniques at least over maize, sunflower, and sugar beet. However, as with any interpretation technique, DL should reach a given level of accuracy, precision, and robustness to satisfy the users' requirements fully.

6.1 Accuracy, precision, and robustness of the models are desired for plant phenotyping

Accuracy is the closeness of the estimated values with the true value. Accuracy is mandatory when estimates are being compared with reference values. It is also desired by breeders if the trait considered is combined with other information or models to get higher-order traits. For example, plant count needs to be accurate if the yield is computed from each plant's grain production. However, when ranking genotypes, precision is mandatory, while accuracy is not. Precision is the closeness between repeated measurements/estimates. Precision quantifies the repeatability, which is essential when identifying differences between genotypes/treatments. Robustness is defined by keeping performances under an extensive range of situations. If estimates of traits need to be precise, robust, and generally accurate, the actual values required by breeders, ecophysiologists and agronomists for these properties are rarely available. The most accepted requirement is to beat a reference method. However, this is difficult to demonstrate in most cases.

Accuracy If the reference method is associated with marginal uncertainties, the reference is good, and any method should be as close as possible to the reference one. It was the case for the plant counting since counting visually on the rows from the images is non-ambiguous and straightforward. However, this was not the case for head counting, where the considered reference is head counting by operators in the field. We demonstrated that the uncertainties associated with this reference method are significant and probably higher than counting on images. In this case, uncertainties of the reference method are due to counting errors in the field by the operator, as well as to the limited spatial representativeness if the whole microplot needs to be characterized. However, if another method is available in addition to the reference and the DL ones, the triple collocation error measurement (TCEM) technique [1] could be used to evaluate the intrinsic accuracy of each method regarding the actual value that is unknown. Further work is to be developed in this direction to quantify better the intrinsic performances of traits estimation techniques

Precision Precision is much easier to quantify using repeatability or broad sense heritability metrics. However, this requires replicates, which was not available in a systematic way in the datasets that we compiled. More attention should be paid to this critical metric, with consequences on selecting the datasets used at least to test the proposed DL models.

Robustness Robustness was the main focus of this work since any machine learning solution

can fail outside its definition domain, i.e. the training domain or "in the domain" (ID). Initial work not reported in this manuscript was trying to characterize the definition domain by describing the space of features extracted from the images of the training dataset. Some results are given in chapter 2 to characterize the datasets added to the Global Wheat Head dataset. However, the results were not as interesting as expected, and the several attempts to optimize the selection of examples to represent the diversity better were deceiving. These techniques were generally describing the distribution of trivial factors associated with the acquisition sessions. However, there were efficient to automatically identify outliers such as blurred images or images corresponding to other scenes than those targeted, such as images of the sky!

One result of our study is that the acquisition session defines well a domain. This is a serious conclusion that implies that most of the literature reported that training and testing over the same acquisition sessions are likely to be not robust. Robustness for phenotyping problems could be defined as the capacity to generalize over all possible sessions. It is measured by the performance drop as proposed in the WILDS study in chapter 4: the change in performances (generally a decrease) when measured over the training domain (ID) and that measured on the test domain (OOD) when the test domain corresponds to acquisition sessions not used for the training. Our work highlights the difficulty to build robust models. This was the case for plant counting: when adding few examples of the test domains, plant identification and counting performances increase drastically. Similar results were reported over the head identification and counting when the "Mixed to Train" setting showed much improved performances compared to the "Train to test setting". It may advocates for active learning, i.e. systematic inclusion of examples from the inference dataset into the training one. However, it is not straightforward to implement for operational applications. Alternative techniques may be implemented to improve model robustness as well as accuracy and precision.

6.2 Possible improvements in the accuracy, precision and robustness of DL approaches

The constitution of large, diverse, and consistent datasets is the necessary first step. Then, data preparation and augmentation, learning strategy, and model design are critical to yield a solution that could be exploited operationally.

6.2.1 Large, diverse and consistent datasets preparation is a crucial first step.

In chapter 1, a dataset composed of 15.000 plants from 3 crops, 18 locations and 27 UAV flights has been gathered across different locations in France. The Global Wheat Head Dataset, described in chapter 2, comprises 6,515 images containing 275,187 wheat heads. These images were collected over 47 acquisition sessions in 16 research institutes across 12 countries. Chapter 2 presented the challenge of building such datasets and how to expose them to the community. The quality and the consistency of the dataset appear very important. The quality is mainly defined by the diversity of the examples, the balance between the several sessions, the quality of the labelling that was reviewed several times in our case, and the quality of the images, avoiding blurriness and under or overexposed images. The consistency across examples is also critical to avoid complexifying the problem for the DL model. The consistency includes a limited range of spatial resolution and view directions and labels clearly defined. The protocol used for image acquisition appears, therefore, mandatory to ensure both quality of images and consistency across sessions. Metainformation is also required to select the images/sessions better and try to understand problems when the model fails on some images/sessions. However, the constitution of large and diverse datasets for training and evaluating robust DL models requires a significant

effort to get images and the associated labels and metainformation from many institutions, check and organize the database, and make it open to the scientific community. This is generally at the initiative of one or a small group of institutions. These databases need to be lively, i.e., to grow and eventually evolve to accept additional traits of interest. For example, after head identification, some characteristics of the heads could be of interest, such as the length, diameter, and a number of spikelets' ranks. This needs to be organized and requires dedicated resources.

Labelling is tedious work to accomplish, although critical and mandatory. It has evolved these last years/months rapidly, with the availability of dedicated ergonomic platforms. Further, the emergence of companies that can be contracted for labelling images for a reasonable price and good quality has changed the paradigm: the remaining tasks to accomplish locally is mostly to precisely define the objects to be labelled and to review the labels proposed by the labelling companies.

6.2.2 Data preparation and augmentation

Data preparation • Illumination is an essential source of variability for RGB imagery as it is often not controlled in a field experiment. Advanced preprocessing techniques can help to mitigate the environmental effect. Relighting of an image from multi-view [2], [3], videos [4] or even single view image [5] has been made possible by recent development in Deep Learning, which could open an avenue to control the illumination. Shadow removal technics are also a possibility to remove illumination [6][7][8][9]. An example of how such technics can perform directly on plant phenotyping imagery is presented in figure 26. Normalization of the pixel values across different RGB cameras can also prevent unexpected behaviour of the Deep Learning algorithms on new cameras or particular illumination conditions.

Figure 26. Presentation of a shadow removal GAN network [7] on wheat head canopy (left: original ; center: GAN ; right : CLAHE).

Data augmentation Before the training, data augmentation generates new examples that provide more diversity. Most studies used such techniques but were limited to geometric (rotation, shear, translation, zoom in and out) that have been exploited to generate new images for seed detections by randomly rearranged individuals who were already labelled [10], [11] with success. However, we demonstrated in chapter 3 that CutMix [12] has a limited impact on the performances, and MixUp[13] may even lower the results. Photometric transforms (random noise, random blur, change of histogram color) may widen the illumination conditions for more comprehensive training of the models. Generating more complex transformations require to use advanced techniques such as plant 3D model [14]–[16] or GAN generated images [2], [3]. As compared to a 2D approach, the 3D one provides more consistent scenes, with realistic interactions between the background and the canopy. Progress on GAN performances with StyleGAN-V2 [17] have been leveraging a new large and fully labelled dataset with

approaches such as [18], allowing to label much more precisely the images – 42 labels for one human face. The recent graphical processing unit also proposes a core dedicated to ray tracing that can generate photo-realistic illuminations, shadows and reflection on a dedicated 3D engine such as Unity. Combining these approaches seems to be interesting for high-resolution synthetic data generation. In the context of plant phenotyping, advanced data augmentation strategies should combine two characteristics:

- take full advantage of the handmade labels, as they are costly and tedious to obtain, and
- include as much expert knowledge as possible to model the diversity of the data.

Geometric transformation and 3D models generate realistic scenes in terms of structure, but the result generally is not photorealistic. In contrast, GAN model can achieve photorealistic texture at the expense of possible alteration of the scene composition structure, which prevents using them directly for a task other than classification. Geometric transformations are easy to implement with no background. Still, when there is a background, some mismatches, such as differences in the illumination of the plants and the soil, result in unrealistic scenes, which cannot be used for training. GAN's success has almost always been achieved for tasks that did not require conserving the structure or only one central object in the image, such as Arabidopsis thaliana in a single pot [16]. Controlling the scene's structure seems essential to include expert knowledge such as the sowing pattern, covering the whole range of the expected phenotype. Generating realistic sets for Deep Learning, as presented in figure 27, would require:

- A process to generate a new and realistic structure of the countable objects on an image.
- A process to generate unique background.
- A process to generate new objects.
- A process to make the background and the objects coherent

The objects in the scene do not need to be generated with a 3D model: actual 3D RGB plants captured in controlled conditions could be used to create a synthetic canopy. The labelling can then be done precisely at the object scale and not at the canopy scale. Synthetic scenes generation pipeline can be re-used for new traits on the same crop. It would access more detailed traits associated to smaller organs such as spikelets and flowers or allow to separate leaves from tillers. Such approaches is likely to increase the number of traits and the level of details that DL can access. The same database of 3D scenes could therefore be re-used to estimate new traits. Such synthetic scenes allow exploring the occlusion problem: the organ can be partially hidden under a leaf, making the measurement difficult. The synthetic scene is crucial for a full perception model that can pre-process RGB images and calculate many measures used to calculate structural traits. It is also a field that will not be explored by another community other than the plant community.

6.2.3 Training

Split of the dataset between training, validation, and test First, the question of "in domain" (ID) and "out of domain" (OOD) is complex and was discussed within the WILDS investigation over several problems, including wheat head detection. For phenotyping applications, a domain may be generally associated with an acquisition session where the sensors, acquisition protocol, illumination conditions, and growth stages are well defined, while genotypes and cultural practices vary. However, a possible bias of our work was the use of data generated from different acquisitions protocols. We expect to be able in the future to define domains beyond the acquisition session, at least when the protocol is controlled. The question of the validation dataset is exciting and was not explicitly discussed in this work. The validation dataset is part of the training process: it is used both for finetuning hyperparameters and

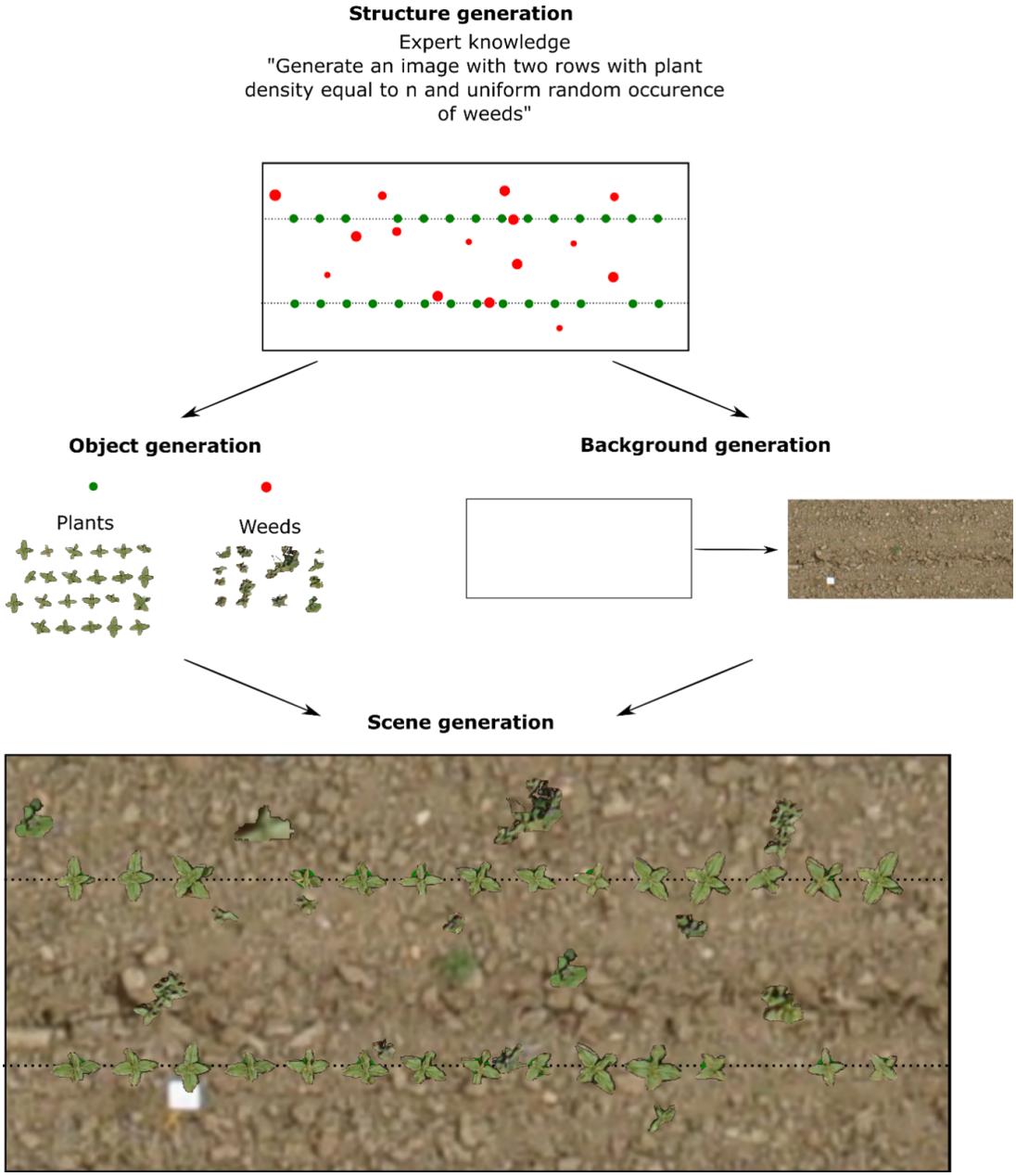


Figure 27. Example of a simple pipeline to generate new UAV microplot from RGB images and previous labels. Expert knowledge can be used to conditioned the generation of new scenes. Objects and background can be retrieve from a bank of image, generate with a 3D model or a GAN

possibly for pseudo-labelling. More emphasis should be put on this vital question: should it be ID or OOD? We recommand for papers presenting DL strategies for traits evaluation, especially for field plant phenotyping, to systematically discuss the domain bias between their train, validation and test splits.

Training algorithm Several algorithms were proposed to improve the robustness of models by playing on the way the coefficients of the network are optimized. In chapter 4, we compared the traditional ERM loss function with more sophisticated algorithms. However, results consistently indicated over the problems and datasets considered in WILDS, that the standard ERM was outperforming these training algorithms. However, their implementation is not trivial, and more work is needed to refine these first results.

Semi-supervision learning In our studies, the dataset used for training always was labelled entirely. However, unlabelled data can help the model generalize over distributions without any labels. This strategy, when a mix of labelled and unlabelled is used to train a DL model, is called semi-supervised learning or sometimes Weakly supervised learning. A straightforward semi-supervision learning strategy is pseudo-labelling which was successfully used by most high-scoring submissions to the Global Wheat Challenges: a model is used to predict labels on a set of unlabelled images. Usually, an ensemble of models is used for more robustness [19], and during training, each batch contains more than 50% of real labelled data. An extension of the GlobalWheat-Wilds is expected to study how to improve robustness with unlabelled data.

Self-supervision learning Self-supervision learning is a branch of Deep Learning which allows training encoders on a large set of unlabeled data points thanks to a surrogate task. These encoders are can directly used for other tasks with transfer learning. The quantity of acquired data is vast within high-throughput phenotyping installations with terabytes of RGB imagery acquired during a growing season. Two families of self-supervised algorithms can be distinguished: (1) contrastive learning aimed at aligning the features of two similar images or two variations of the same images [20], [21]; (2) predictive approach, for instance, with solving a jigsaw [22] or by predicting a class determined with a simple clustering approach on the deep features (DeepCluster V2 [23], Swav [24]). Self-supervision learning is also becoming more accessible thanks to open-source libraries such as Facebook VISSL [25]. Another specific aspect of the phenotyping platform is to have a fixed set of sensors that make simultaneous observations of the same scene portion. Multimodal Deep Learning aims at learning a representation not only for RGB images but also for Lidar or multispectral images and is already used for autonomous driving [26]. Self-supervision learning is also applied to the multimodal model: OpenAI recently proposed CLIP [27], a model trained jointly on text and images. A consequence of Selfsupervision Learning is the emergence of huge models with billions of parameters such as GPT-3 [28], also called "foundation models" [29]. A characteristic of these very large models is the ability to execute new tasks without any training (zero-shot learning). However, they are challenging to train as it requires massive computing power and can be viewed as an infrastructure more than an algorithm.

Multitask learning

One DL algorithm is generally designed to access one trait or provide a pre-processing of the raw image. With a growing number of traits to measure from the same images, running multiples DL models in parallel can be burdensome in practice. Additionally, each new trait will generate a different annotated dataset that is not used for another task, which can be seen as underuse of the data. As described in the introduction, most DL models will learn to extract meaningful features from the data before solving the problem. The same features could be learned from more extensive datasets for multiple tasks. Multitask models can learn better representation from several data sources and have already been explored for wheat phenotyping [30] and Arabidopsis phenotyping [31]. It was shown that it requires fewer labels for underrepresented tasks and reduces overfitting [32], while improving adversarial robustness [33]. A comparison of both approaches is presented in figure 28. Learning an encoder for all tasks of one platform instead of having different models allows one to use all the available data and avoid overfitting if only small datasets are available. The learned representation can also be used for other purposes than estimating traits. It could be used to monitor outliers or discover subgroups.

6.2.4 Model selection

The baseline model was selected after the pioneering study of Madec et al. 2019 [34]. It is based on the Faster-RCNN [35] that first detect potential objects of interest and then

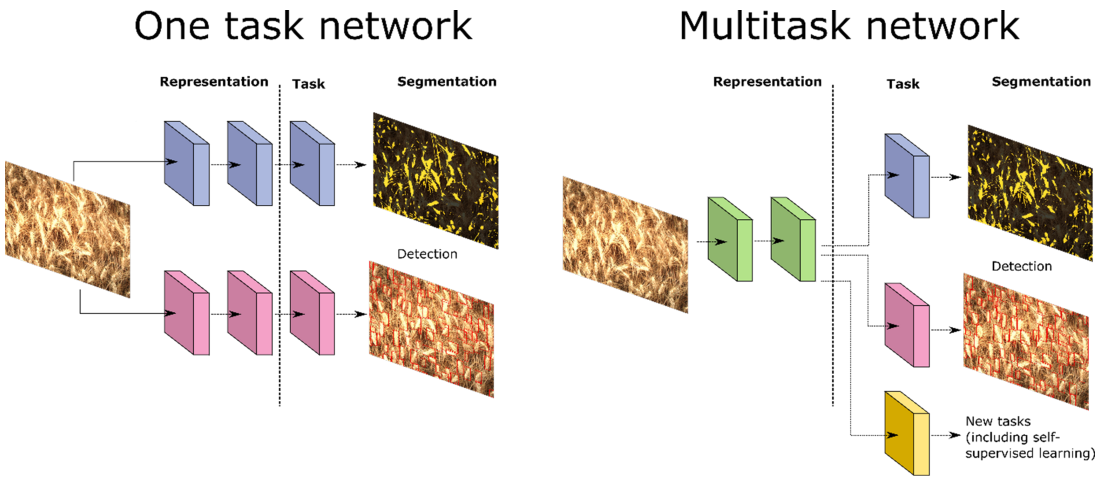


Figure 28. Comparison of a system based on multiple networks with separated encoder ("one task network") and tasks networks sharing the same encoder which is improved with all labels ("Multitask network").

find out which should be considered the actual ones. The challenges organized offered the opportunity to evaluate the diversity of models used and their associated performances. Results demonstrated that the standard one-step models such as YoloV3 [36] and V5 and [37] were part of the winning solutions proposed. However, it is not easy to evaluate the contribution of the architecture of the model used to the observed performances of the solution. More investigation is needed to quantify the contribution of each element of the solution.

6.3 The integration of Deep Learning to estimate plant traits at Arvalis

The doctoral thesis results were obtained in the context of a collaboration between the National Research Institute for Agriculture and Environment (Institut National de Recherche en Agronomie et Environnement – INRAE), and Arvalis a private research institute funded by the French farmers. Its mission is to disseminate innovations in agriculture to farmers by conducting research and evaluating novelties for more than 60 years. Arvalis is deploying its high-throughput phenotyping technologies for its use on various trial stations throughout France, ranging from affordable and low-tech machines like LITERAL to high-tech, automated systems like Phenomobile and UAVs. Robust estimates of plant traits are then a requirement to put in production technologies based on Deep Learning. More challenges arise with production problems: the code needs to be fast enough to process several sessions of acquisitions for one night, and the results need to be appropriately tested against manual baseline. Arvalis provided all the required manual and automatic data during three campaigns for the study. In return, the codes developed for the papers were rewritten adequately for production usage, with a specific focus on the clarity of code and the execution speed and tested within the phenotyping team. I also work a lot on the new architecture of the whole processing chain and the processing code of several other Deep Learning modules, such as semantic segmentation. For instance, the plant counting pipeline developed for Chapter 1 and the wheat head counting pipeline for Chapter 3 are today used. In 2021, Arvalis conducted several non-methodological trials without manual measurements of the wheat head density, relying on the Deep Learning wheat head density estimation module.

6.4 References

[1] A. Stoffelen, "Toward the true near-surface wind speed: Error modeling and calibration using triple collocation," *Journal of Geophysical Research: Oceans*, vol. 103, no. C4, pp. 7755–7766, 1998. DOI: https://doi.org/10.1029/97JC03180. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/97JC03180. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/97JC03180.

[2] J. Philip, M. Gharbi, T. Zhou, A. A. Efros, and G. Drettakis, "Multi-view relighting using a geometry-aware network," en, *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1–14, Jul. 12, 2019, ISSN: 0730-0301, 1557-7368. DOI: 10.1145/3306346.3323013.

[3] *Deep image-based relighting from optimal sparse samples*. [Online]. Available: https://cseweb.ucsd.edu/%7B%5Ctextasciitilde%7Dviscomp/projects/SIG18Relighting/.

[4] X. Luo, J.-B. Huang, R. Szeliski, K. Matzen, and J. Kopf, "Consistent video depth estimation," *arXiv:2004.15021 [cs]*, Aug. 26, 2020. [Online]. Available: http://arxiv.org/abs/2004.15021.

[5] J. Kopf, K. Matzen, S. Alsisan, O. Quigley, F. Ge, Y. Chong, J. Patterson, J.-M. Frahm, S. Wu, M. Yu, P. Zhang, Z. He, P. Vajda, A. Saraf, and M. Cohen, "One shot 3d photography," *arXiv:2008.12298 [cs]*, Sep. 1, 2020. [Online]. Available: http://arxiv.org/abs/2008.12298.

[6] H. Cheong, S. K. Devalla, T. H. Pham, L. Zhang, T. A. Tun, X. Wang, S. Perera, L. Schmetterer, T. Aung, C. Boote, A. Thiery, and M. J. A. Girard, "Deshadowgan: A deep learning approach to remove shadows from optical coherence tomography images," en, *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 23–23, Jan. 28, 2020, ISSN: 2164-2591. DOI: 10.1167/tvst.9.2.23.

[7] X. Hu, Y. Jiang, C.-W. Fu, and P.-A. Heng, "Mask-shadowgan: Learning to remove shadows from unpaired data," en, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South): IEEE, Oct. 2019, pp. 2472–2481, ISBN: 978-1-72814-803-8. DOI: 10.1109/ICCV.2019.00256. [Online]. Available: https://ieeexplore.ieee.org/document/9010942/.

[8] H. Le and D. Samaras, "Physics-based shadow image decomposition for shadow removal," *arXiv:2012.13018 [cs]*, Dec. 23, 2020. [Online]. Available: http://arxiv.org/abs/2012.13018.

[9] X. Cun, C.-M. Pun, and C. Shi, "Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan," *arXiv:1911.08718 [cs]*, Nov. 20, 2019. [Online]. Available: http://arxiv.org/abs/1911.08718.

[10] D. Ward, D. Ward, P. Moghadam, P. Moghadam, N. Hudson, and N. Hudson, "Deep leaf segmentation using synthetic data," en, p. 13,

[11] D. Ward and P. Moghadam, "Scalable learning for bridging the species gap in image-based plant phenotyping," *arXiv:2003.10757 [cs]*, Apr. 23, 2020. [Online]. Available: http://arxiv.org/abs/2003.10757.

[12] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features," en, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 6022–6031, ISBN: 978-1-72814-803-8. DOI: 10.1109/ICCV.2019.00612. [Online]. Available: https://ieeexplore.ieee.org/document/9008296/ (visited on 09/23/2021).

[13] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *CoRR*, vol. abs/1710.09412, 2017. arXiv: 1710.09412. [Online]. Available: http://arxiv.org/abs/1710.09412.

[14] E. Jacopin, N. Berda, L. Courteille, W. Grison, L. Mathieu, A. Cornuéjols, and C. Martin, "Using agents and unsupervised learning for counting objects in images with spatial organization," SciTePress, 2021, pp. 688–697, ISBN: 978-989-758-484-8. DOI: 10.5220/0010228706880697.

[15] W. Zhang, K. Chen, J. Wang, Y. Shi, and W. Guo, "Easy domain adaptation method for filling the species gap in deep learning-based fruit detection," en, *Horticulture Research*, vol. 8, no. 1, pp. 1–13, Jun. 1, 2021, ISSN: 2052-7276. DOI: 10.1038/s41438-021-00553-8.

[16] M. V. Giuffrida, H. Scharr, and S. A. Tsaftaris, "Arigan: Synthetic arabidopsis plants using generative adversarial network," 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Oct. 2017, pp. 2064–2071. DOI: 10.1109/ICCVW.2017.242.

[17] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *CoRR*, vol. abs/1912.04958, 2019. arXiv: 1912.04958. [Online]. Available: http://arxiv.org/abs/1912.04958.

[18] Y. Zhang, H. Ling, J. Gao, K. Yin, J.-F. Lafleche, A. Barriuso, A. Torralba, and S. Fidler, "Datasetgan: Efficient labeled data factory with minimal human effort," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, Computer Vision Foundation / IEEE, 2021, pp. 10 145–10 155.

[19] I. Radosavovic, P. Dollár, R. B. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," *CoRR*, vol. abs/1712.04440, 2017. arXiv: 1712.04440. [Online]. Available: http://arxiv.org/abs/1712.04440.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," *CoRR*, vol. abs/2002.05709, 2020. arXiv: 2002.05709. [Online]. Available: https://arxiv.org/abs/2002.05709.

[21] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," en, *arXiv:1911.05722 [cs]*, Mar. 23, 2020. [Online]. Available: http://arxiv.org/abs/1911.05722.

[22] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., ser. Lecture Notes in Computer Science, vol. 9910, Springer, 2016, pp. 69–84. DOI: 10.1007/978-3-319-46466-4_5.

[23] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," en, *arXiv:2006.09882 [cs]*, Jan. 8, 2021. [Online]. Available: http://arxiv.org/abs/2006.09882.

[24] P. Goyal, M. Caron, B. Lefaudeux, M. Xu, P. Wang, V. Pai, M. Singh, V. Liptchinsky, I. Misra, A. Joulin, and P. Bojanowski, "Self-supervised pretraining of visual features in the wild," en, *arXiv:2103.01988 [cs]*, Mar. 5, 2021. [Online]. Available: http://arxiv.org/abs/2103.01988.

[25] P. Goyal, Q. Duval, J. Reizenstein, M. Leavitt, M. Xu, B. Lefaudeux, M. Singh, V. Reis, M. Caron, P. Bojanowski, A. Joulin, and I. Misra, *Vissl*, https://github.com/facebookresearch/vissl, 2021.

[26] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López, "Multimodal end-to-end autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2020, ISSN: 1524-9050, 1558-0016. DOI: 10.1109/TITS.2020.3013234.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *arXiv:2103.00020 [cs]*, Feb. 26, 2021. [Online]. Available: http://arxiv.org/abs/2103.00020.

[28] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *arXiv:2005.14165 [cs]*, Jul. 22, 2020. [Online]. Available: http://arxiv.org/abs/2005.14165.

[29] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Kohd, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the opportunities and risks of foundation models," *arXiv:2108.07258 [cs]*, Aug. 18, 2021. [Online]. Available: http://arxiv.org/abs/2108.07258.

[30] M. P. Pound, J. A. Atkinson, D. M. Wells, T. P. Pridmore, and A. P. French, "Deep learning for multi-task plant phenotyping," 2017, pp. 2055–2063.

[31] A. Dobrescu, M. V. Giuffrida, and S. A. Tsaftaris, "Doing more with less: A multitask deep learning approach in plant phenotyping," English, *Frontiers in Plant Science*, vol. 11, 2020, ISSN: 1664-462X. DOI: 10.3389/fpls.2020.00141. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpls.2020.00141/full.

[32] J. Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," *Machine learning*, vol. 28, no. 1, pp. 7–39, 1997.

[33] C. Mao, A. Gupta, V. Nitin, B. Ray, S. Song, J. Yang, and C. Vondrick, "Multitask learning strengthens adversarial robustness," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, Springer, 2020, pp. 158–174.

[34] 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," en, *Agricultural and Forest Meteorology*, vol. 264, pp. 225–234, Jan. 15, 2019, ISSN: 0168-1923. DOI: 10.1016/j.agrformet.2018.10.013.

[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2015, pp. 91–99.

[36] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[37] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, Computer Vision Foundation / IEEE, 2020, pp. 10 778–10 787.