On the Design of the Agri-Food Competition for Robot Evaluation (ACRE)

Riccardo Bertoglio*, Giulio Fontana*, Matteo Matteucci*, Davide Facchinetti†, Michel Berducat‡, and Daniel Boffety‡

*Politecnico di Milano - DEIB, Milan, Italy
Email: {riccardo.bertoglio,giulio.fontana,matteo.matteucci}@polimi.it

†Università degli Studi di Milano - DISAA, Milan, Italy
Email: davide.facchinetti@unimi.it

‡INRAE Clermont-Ferrand, Montoldre, France
Email:{daniel.boffety,michel.berducat}@inrae.fr

***Abstract*—The Agri-Food Competition for Robot Evaluation (ACRE) is a novel competition for autonomous robots and smart implements. It is focused on agricultural tasks such as removing weeds or mapping/surveying crops down to single-plant resolution. Such abilities are crucial for the transition to so-called "Agriculture 4.0", i.e., precision agriculture supported by ICT, Artificial Intelligence, and Robotics. ACRE is a *benchmarking competition*, i.e., the activities that participants are required to execute are structured as performance benchmarks. The benchmarks are grounded on the key scientific concepts of objective evaluation, repeatability, and reproducibility. Transferring such concepts in the agricultural context, where large parts of the test environment are not fully controllable, is one of the challenges tackled by ACRE. The ACRE competition involves both physical Field Campaigns and data-based Cascade Campaigns. In this paper, we present the benchmarks designed for both kinds of Campaigns and report the outcome of the ACRE dry-runs that took place in 2020.**

***Index Terms*—robotics, smart agriculture, benchmarking, competition, machine learning, image segmentation**

I. Introduction

METRICS [1] is an EU-funded project dedicated to the metrological evaluation and testing of autonomous robots. One of the key activities of METRICS is the organization of four robot benchmarking competitions. ACRE (Agri-food Competition for Robot Evaluation) [2] is one of these competitions and deals with the applications of robotics to agriculture. ACRE (as the other METRICS competitions) is based on the concept of "benchmarking through competitions" [3], i.e., on exploiting the appeal and desirable features of competitions to foster a culture of benchmarking in European robotics. In this, ACRE shares the approach with other ongoing efforts, such as the European Robotics League [4]–[6].

Robotics competitions organized through the years are many. The most notable include the EUROBENCH project [7] focused on bipedal robotics technologies (exoskeletons, prostheses, and humanoids); the RoboCup, RoCKIn and European Robotics League [8] focused on indoor competitions related to domestic and industrial robots; and the euRathlon [9], the world's first multi-domain (air, land and sea) multi-robot search and rescue competition. Regarding the agricultural domain, it has to be mentioned the Tomato Harvesting Robot Competition [10] focused on the evaluation of robots for automated tomato harvesting to reduce the working time of harvesting. However, to the best knowledge of this paper's authors, no agricultural weeding robot competitions have been held in the past, apart from the ROSE Challenge on which ACRE builds on.

The ROSE Challenge [11] is a French national project ending in 2021 involving companies developing weeding robots. The ROSE Challenge is organized by the National Laboratory of Metrology and Testing (LNE) and the National Research Institute for Agriculture, Food and the Environment (INRAE). The ROSE challenge aims to encourage the development of innovative solutions for intra-row weed control to reduce or even eliminate herbicides. Four teams are competing against each other during the challenge.

A specific challenge for ACRE comes from the fact that its Field Campaigns occur in outdoor environments and that a key element of the experimental setup is live crops. Consequently, a longer and less controllable preparation phase is required due to plant growth. Moreover, like the actual state of readiness at the time of the Campaign, the setup's quality heavily depends on the weather. Indeed, some benchmarks (see Section II) require plants at a specific growth stage.

ACRE competition comprises two separate but interconnected tracks. One track comprises the so-called *Field Campaigns* that involve robots executing activities in real-world agricultural environments such as open-air fields. The events of the second track are called *Cascade Campaigns*. In Cascade Campaigns, Artificial Intelligence systems perform activities on data collected during the Field Campaigns. Data are like crop images or robot sensor logs.

There is mutual feedback between the two tracks of ACRE; indeed, Field Campaigns provide data to Cascade Campaigns,

This project has received funding from the European Commission under H2020 grant agreement No 871252 "METRICS: Metrological Evaluation and Testings of Robots in International Competitions".

while Cascade Campaigns are a way to foster the development of systems to be exploited in subsequent Field Campaigns. Cascade Campaigns are a way to involve in ACRE people experienced in data analytics without necessarily acquiring robotics expertise. At the same time, they can also present their achievements to interested parties, such as companies involved in agricultural robotics, which lack expertise in perception.

The first events in the Field and Cascade tracks of ACRE took place in 2020, respectively, in October 2020 and from October 2020 to January 2021. We officially called these events "dry-runs" because we used them to verify and validate the organization more than to provide rankings representing the state-of-the-art. The next ACRE Field and Cascade Campaigns are foreseen for June 2021 and October 2021, respectively.

In the next section, a description of the ACRE benchmarks is provided. Sections III and IV describe, respectively, the results of the dry-run Field Campaign and Cascade Campaign. Section V presents the outlook for future editions of the competition and a brief discussion on the pedagogical impact of Cascade Campaigns.

II. Overview of ACRE Benchmarks

ACRE's benchmarks take two forms [12]: *Functionality Benchmarks (FBMs)*, focused on specific capabilities of a robot and designed to make the benchmark as independent as possible from robot components not directly involved in the functionality under examination (FBMs are *Plant discrimination*, *Field navigation*, *Leaf area estimation*, *Weed destruction*, *Biomass estimation*); *Task Benchmarks (TBMs)*, evaluating the execution of complex tasks involving multiple functionalities, where the final result depends both on the individual functionalities and on the integration between components (TBMs are *Intra-row weeding*, and *Crop mapping*). We have described ACRE benchmarks with detailed information on their execution and evaluation metrics in the ACRE Evaluation Plan [13]. Below we provide an overview of them.

ACRE benchmarks involve three different robot capabilities required in agricultural tasks: robot perception, navigation, and manipulation. ACRE benchmarks concerning perception are:

- **Plant discrimination (FBM)** evaluates the capability of discriminating which plants of a row are weeds and which are crops (intra-row detection). The robot is required to make a pass over a prepared row containing both crops and weed plants; using its sensors (e.g., vision), the robot classifies the crops and weeds present in the rows. To decouple plant discrimination functionality from others, we do not require the robot to move autonomously.
- **Leaf area estimation (FBM)** evaluates the capability of estimating the plants' leaf area along a cultivated row. The test environment is a linear row with plant height in the range $30\,\mathrm{cm}$–$50\,\mathrm{cm}$. The robot must move along the row and use its perception to estimate the variable leaf area along the entire row. We do not require the robot to move autonomously.
- **Biomass estimation (FBM)** evaluates the capability of estimating above-ground crop biomass. The robot must make a pass over a prepared field composed of one or more rows, using its sensors to perceive the plants. The robot must estimate the fresh weight of the above-ground parts of the plants (without distinguishing between types of plants). We do not require that robots move autonomously.

ACRE benchmarks concerning robot navigation are:

- **Field Navigation (FBM)** requires the robot to move through cultivation without damaging the crop. The organizers identify predefined destination locations within a cultivated area. The robot under test is assigned one of these locations and required to reach it within a timeout.
- **Crop Mapping (TBM)** evaluates the robot's capability to produce a map of the entire cultivation by exploring it autonomously. The robot must explore a multi-row cultivated plot autonomously and provide a map of crop plants. The robot has to recognize single plants and provide their positions on a Cartesian coordinate system.

ACRE benchmarks concerning manipulation, in the wider sense of "physical interaction with the environment", are:

- **Weed Destruction (FBM)** evaluates the capability of destroying weeds in intra-row without damaging crops. The evaluation compares the state of the test plot before and after the weeding operations. To make this evaluation as independent as possible from other functionalities, we use visual markers to identify crop and weed plants in the prepared plot; additionally, the robot is not required to drive autonomously along the row.
- **Intra-Row Weeding (TBM)** assesses the capability to perform fully autonomous intra-row weeding of a row. The robot must eliminate the weeds located among a row's crop plants without damaging the crop. The robot must navigate the rows autonomously, and there are no markers on the plants to facilitate their detection and identification.

The benchmarks described above refer to the Field Campaigns, and only a subset of them will be executed in every Campaign based on the participants' implemented capabilities. The organizers are currently selecting the benchmarks for the next (2021) Campaign together with interested stakeholders. Instead, since ACRE Cascade Campaigns do not involve physical robots, we cannot implement all of the ACRE benchmarks.

In Cascade Campaigns, the set of possible benchmarks is limited to those concerning pure perception (i.e., Plant discrimination, Leaf area estimation, Biomass estimation). An additional benchmark is a modified version of the Crop Mapping TBM where the robot's trajectory has been recorded in a previous Field Campaign and thus cannot be controlled.

III. ACRE Dry-Run Field Campaign

The ACRE dry-run Field Campaign took place in October 2020 at the experimental farm of the National Research Institute for Agriculture, Food and the Environment (INRAE)

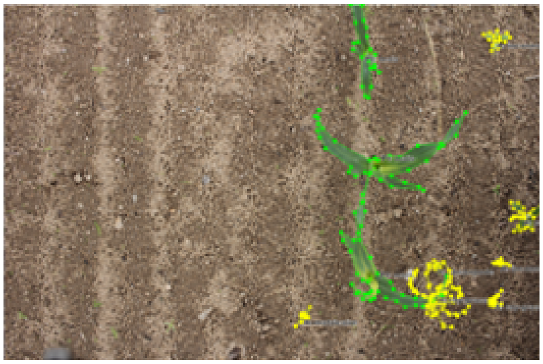


Fig. 1: Example of ground truth image labeled by a human expert; colored dots outline the areas identified as weeds (yellow) or crops (green).

located in Montoldre (France). We exploited the dry-run event to test and adjust the benchmarking protocols according to the METRICS project methodology [14], and to obtain datasets for the future ACRE Cascade Campaign. We made to coincide the ACRE dry-run in time and space with the ROSE Challenge. We aimed to exploit the synergy between the ACRE and ROSE challenges and encourage ROSE teams to participate in both.

The dry-run setup has involved multiple benchmarks, i.e., Plant Discrimination FBM, Weed Destruction FBM, Intra-Row Weeding TBM, and Field Navigation FBM. The first three of these have been conducted in parallel with the second evaluation campaign of the ROSE challenge since the experimental setup was designed to be compatible. Due to the circumstances of the COVID-19 pandemic, the fourth benchmark (Field Navigation FBM) was executed by only one team.

A. Preparation of Plant Discrimination, Weed Destruction, and Intra-Row Weeding benchmarks

Plant Discrimination requires that a robot uses its perception and interpretation capabilities to differentiate between crops and weeds based on their features. The plant classification produced by the robot can be compared with ground truth provided by qualified humans. Figure 1 shows an example of ground truth provided by human experts. Participants are evaluated based on the EGER metric [15] that accounts for the number of correctly and incorrectly classified and missed plants.

In the Weed Destruction FBM, the goal is to destroy weeds without damaging the crops. To decouple this capability from the robot's performance in discriminating weeds from crops, we labeled plants using colored discs as shown in Figure 2. Since Intra-Row Weeding is a Task Benchmark, no visual markers are allowed for this benchmark, and the robot must eliminate the weeds autonomously. Participants are evaluated by manually counting the number of destroyed weeds and damaged plants (see Figures 3, and 4).

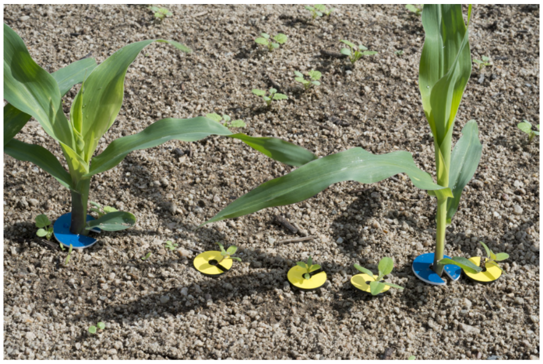


Fig. 2: Crops and weeds identified with colored discs for the Weed Destruction FBM.

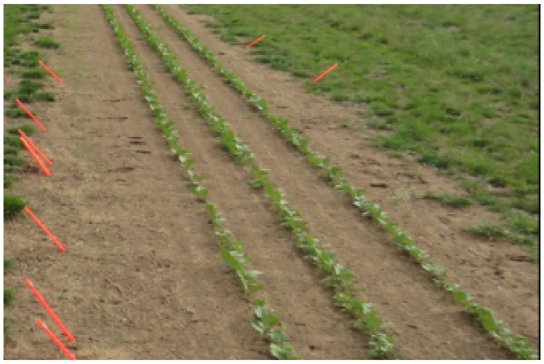


Fig. 3: Rows prepared for the Intra-Row Weeding Benchmark.

B. Preparation of the Field Navigation FBM

The test field used for the dry-run of the Field Navigation FBM was divided into two four experimental plots of $2\,\mathrm{m}$ wide by $46.5\,\mathrm{m}$ long that have been prepared and sown with two maize crop rows each. The inter-row spacing was $75\,\mathrm{cm}$, and the mean maize plant spacing on the sowing line was $14\,\mathrm{cm}$. We sowed two maize plots with straight lines of plants and the other two with a curve in the middle of the row by shifting the lines by an offset. At each end of the plot, a free grassy area allowed robots to realize the half-turn required to invert their motion direction. In the two plots with the offset, maize rows were straight for about $10\,\mathrm{m}$ before and after the offset.

Due to the dry-run meteorological conditions, maize growth was limited. The maize plants reached only the growth stage in which they show two or three small yellow leaves for an approximate height between five and six centimetres. Figure 6 shows the configuration of the four maize plots used for the FBM.

C. Execution of the Field Navigation FBM in the dry-run

Due to the ongoing pandemic's strong constraints, only one participant executed the Field Navigation FBM: namely SITIA, a French company. SITIA, a partner of the Roseau team in the ROSE challenge, participated with its Trektor platform (see Figure 5). Trektor is a hybrid platform developed for agricultural applications of market gardening and viticulture.

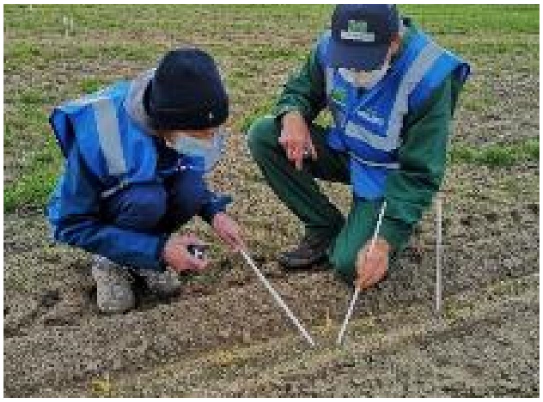


Fig. 4: Assessment of the damaged crop plants and counting of destroyed weeds in the Intra-Row Weeding Benchmark.

SITIA's execution of ACRE's FBM took place on October 22th, 2020.

In preparation for the Field Navigation FBM, a SITIA operator recorded several points between the maize rows on each plot with a portable RTK GPS. The recorded trajectories allowed the Trektor to perform the field navigation FBM later. It must be noted that pre-recording waypoints, while perfectly acceptable in the dry-run context, are not compatible with the rules of the ACRE competition, which require that the robots exploit onboard perception to follow the plant rows.

The Trektor platform, placed manually in front of the first straight row at the starting point, successfully followed the maize rows until the end using the pre-recorded GPS points (see Figure 5). All trajectories were executed without damaging the crop, as required by the benchmark. At each end of the maize plot, the SITIA operator manually controlled the half-return required to invert motion direction. Again, it must be noted that manual driving is compatible with the dry-run but not with the full rules of the Field Navigation FBM.

IV. ACRE Dry-Run Cascade Campaign

In the ACRE Dry-Run Cascade Campaign, we built upon the data collected by the participants of the 2019 ROSE Challenge. We set up an online competition asking to segment RGB images to distinguish between crop, weeds, and background. Automatic crop and weed segmentation can be a driver of innovations to optimize the agricultural processes. A ground robot can exploit automatic weed detection for mechanical weeding; thus, the use of chemicals could even be avoided entirely.

The competition was published via CodaLab Competitions [16], a powerful open-source framework for running competitions that involve results or code submission. Organizers can program many aspects of the competition, thus having more space for customization. It is also possible to run the competitions in the organizers' docker, and their compute workers. Since we did not need a high computing workforce, we hosted our competition on the CodaLab servers.

In the competition, we had 57 teams accounting for 457 individuals. On the total number of participants, 95.6% were from 53 different institutions (universities, research centers, and companies). The remaining 4.4% declared to be not attached to any institution for the scope of the competition.

The competition duration was 97 days, from 17 October 2020 to 22 January 2021 (plus one day of extension on 29 January 2021). It has been divided into two phases, *Development* and *Final*. In the Development phase, participants were required to train their models on the Training set and submit predictions of the Test_Dev set. At the end of the Development stage, we released the labels of the Test_Dev set and the new, unseen and unlabeled, Test set. Thus, in the Final phase, participants were required to submit predictions of the new Test set. The Final phase was restricted to the last three days of the competition and characterized by a limited number of submissions (a maximum of 10 submissions). The limit to the number of submissions was imposed to reduce the risk of overfitting.

The dataset was composed of images captured by different sensors in different moments and was about two kinds of crops: haricot and maize. Data came from the 2019 ROSE Challenge, where four teams have competed with agricultural robots. The names of these four teams are Bipbip, Pead, Roseau, and WeedElec. Each team has collected images of the same two crops, but in different moments and with different sensors (RGB cameras). The dataset contained both RGB images and some labeled masks (ground truth). Masks were composed of three different classes: crop, weed, and background. Figure 7 shows an example of an RGB image and its corresponding labeled mask. Dataset images were divided by the team that acquired the image, and for each team, by the type of crop present in the images, i.e., haricot and maize. In particular, the dataset was composed as in the following:

- 90 Training images (per team per crop)
- 15 Test_Dev images (per team per crop)
- 20 Test images (per team per crop)

Thus, since we had four teams and two types of crops, the total number of images in the dataset was 1000.

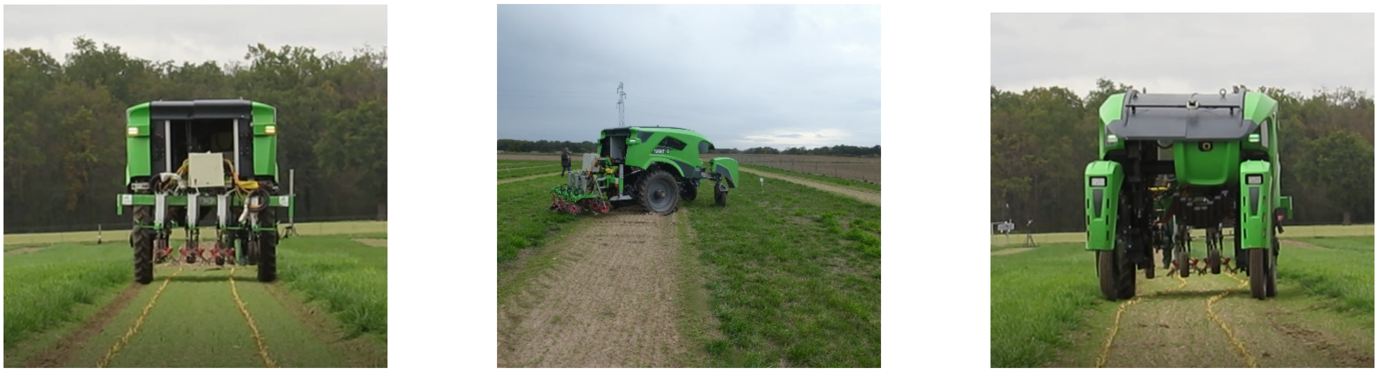
Participants were evaluated on the mean Intersection over Union (IoU) obtained on the two classes, crop, and weed. The Intersection over Union, also called Jaccard Index [17], is typically used in segmentation tasks, and it essentially quantifies the percentage of overlap between predicted and target segmentations. If A is the prediction and B is the ground truth, the IoU is calculated as in the following:

$$IoU = \frac{\mid A \cap B \mid}{\mid A \cup B \mid}.$$

IoU was computed for each target class (crop and weed) separately, by considering prediction and ground truth as binary masks. Then, the final IoU is computed by averaging the two. Thus, we had the following formulation:

$$IoU_{crop} = TP_{crop}/(TP_{crop} + FP_{crop} + FN_{crop})$$

$$IoU_{weed} = TP_{weed}/(TP_{weed} + FP_{weed} + FN_{weed})$$



(a) First pass (straight row).

(b) Manually controlled half-turn.

(c) Second pass (curved row).

Fig. 5: The Trektor robot executing the Navigation FBM during the ACRE dry-run Field Campaign.

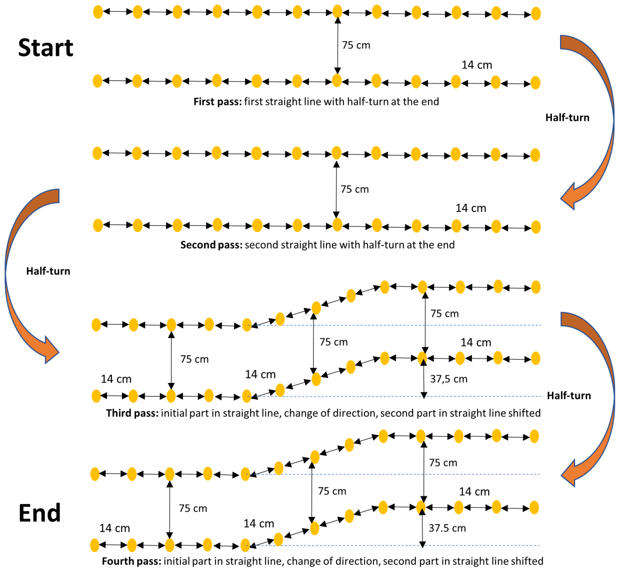

Start
75 cm
14 cm
First pass: first straight line with half-turn at the end
Half-turn
75 cm
14 cm
Second pass: second straight line with half-turn at the end
Half-turn
75 cm
75 cm
14 cm
14 cm
14 cm
37,5 cm
Half-turn
Third pass: initial part in straight line, change of direction, second part in straight line shifted
75 cm
75 cm
14 cm
End
75 cm
14 cm
14 cm
37.5 cm
Fourth pass: initial part in straight line, change of direction, second part in straight line shifted

Fig. 6: Robot path for the Navigation FBM. In the two bottom lines you can note the central offset.

$$IoU = (IoU_{crop} + IoU_{weed})/2$$

where TP are the True Positives, FP are the False Positives and FN are the False Negatives.

Thanks to the CodaLab Competitions framework's flexibility, we could score the participants with different customized IoUs. In particular, we scored the participants according to the "Global IoU" (by considering the images of both crops and the four teams), the Haricot and Maize IoUs, and an IoU for each of the four teams (Bipbip, Pead, Roseau, and WeedElec). Thus, we nominated seven competition winners for each of the categories above.

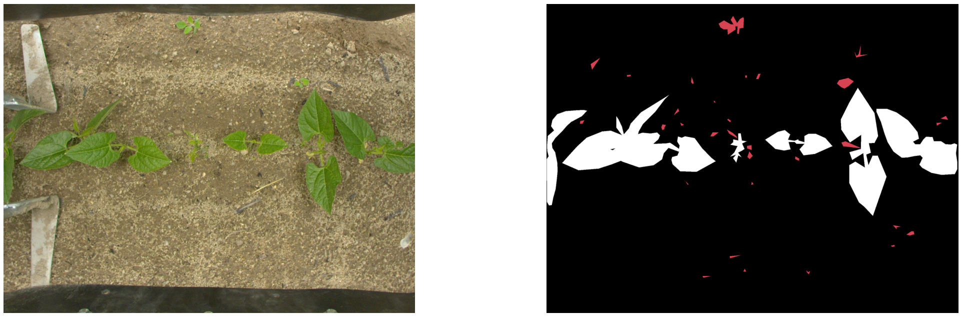
Figure 8 shows the daily-wise distribution of the total number of submissions and the evolution of the daily highest score (Global IoU). The highest Global IoU was 0.7858 in the Development phase and 0.7753 in the Final phase. In the Development phase, the highest IoU related to teams Bipbip, Pead, Roseau, and WeedElec, was 0.8272, 0.6256, 0.7129, and 0.8319, respectively. In the Final phase, the highest IoU related to teams Bipbip, Pead, Roseau, and WeedElec, was 0.8189, 0.6483, 0.7359, and 0.8115, respectively. In the Development phase, the highest IoU related to Haricot and Maize was 0.7954 and 0.7720. In the Final phase, the highest IoU related to Haricot and Maize was 0.7740 and 0.7748. The results obtained in the Final phase were, in general, lower but still close to those of the Development phase. The team Pead images were the most challenging to segment in both phases, probably due to the RGB camera's different positioning. Haricot and Maize images did not show relevant differences in the corresponding IoU scores, suggesting a similar complexity of the task.

V. Conclusions

In this paper, we have presented the design of a robotics competition in the agricultural setting. The competition has been designed around the complex task of intra-row weeding to develop autonomous robots capable of pesticide-free weed destruction. This task requires robot perception, navigation, and manipulation capabilities. These capabilities are evaluated in an objective, repeatable, and reproducible way via an established methodology that is based on task and functionality benchmarks [12].

The design of outdoor agricultural competitions poses several challenges. For instance, weather conditions' uncertainty has a clear impact on the stage of crops at the competition time. To deal with this uncertainty, we organized the first edition of the ACRE competition in dry-run mode. Indeed, we aimed to investigate the benchmarks' feasibility in the setting where the Field Campaigns will happen in the following years.

As a follow-up of Field Campaigns, ACRE foresees Cascade Campaign based on data acquired on the field. Cascade Campaigns are aimed at targeting the Artificial Intelligence community with tasks such as Plant Discrimination. The first Cascade Competition has seen many participants as it did not require a physical robot but just software components. It is worth noticing that the online competition has attracted companies' and universities' involvement, being the latter also



(a) RGB image

(b) Labeled mask

Fig. 7: A couple of RGB image and corresponding ground truth mask from the dataset used in the 2020 ACRE dry-run Cascade Campaign.

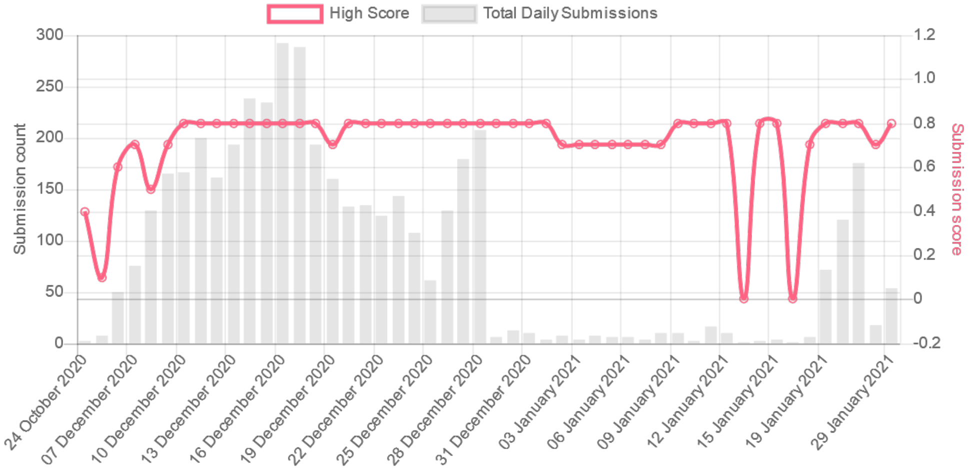


Fig. 8: Daily-wise distribution of the total number of submissions and evolution of the daily highest score (Global IoU).

interested in using the competition data to have students facing real-world problems in their AI and Machine Learning courses.

REFERENCES

[1] METRICS: Metrological evaluation and testing of robots in international competitions. [Online]. Available: https://metricsproject.eu/

[2] ACRE: Agri-food competition for robot evaluation. [Online]. Available: https://metricsproject.eu/agri-food/

[3] F. Amigoni, A. Bonarini, G. Fontana, M. Matteucci, and V. Schiaffonati, "To what extent are competitions experiments? a critical view," in *Workshop on Epistemological Issues in Robotics Research and Research Result Evaluation; Hong Kong. ICRA*, vol. 5, 2014.

[4] ERL: the european robotics league. [Online]. Available: https://www.eu-robotics.net/robotics_league/

[5] F. Ferreira, G. Ferri, Y. Petillot, X. Liu, M. P. Franco, M. Matteucci, F. J. P. Grau, and A. F. Winfield, "Scoring robotic competitions: Balancing judging promptness and meaningful performance evaluation," in *2018 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. IEEE, 2018, pp. 179–185.

[6] M. Basiri, E. Piazza, M. Matteucci, and P. Lima, "Benchmarking functionalities of domestic service robots through scientific competitions," *KI-Künstliche Intelligenz*, vol. 33, no. 4, pp. 357–367, 2019.

[7] D. Torricelli and J. L. Pons, "Eurobench: Preparing robots for the real world," in *International Symposium on Wearable Robotics*. Springer, 2018, pp. 375–378.

[8] P. U. Lima, D. Nardi, G. K. Kraetzschmar, R. Bischoff, and M. Matteucci, "Rockin and the european robotics league: building on robocup best practices to promote robot competitions in europe," in *Robot World Cup*. Springer, 2016, pp. 181–192.

[9] A. F. Winfield, M. P. Franco, B. Brueggemann, A. Castro, M. C. Limon, G. Ferri, F. Ferreira, X. Liu, Y. Petillot, J. Roning *et al.*, "eurathlon 2015: A multi-domain multi-robot grand challenge for search and rescue robots," in *Annual Conference Towards Autonomous Robotic Systems*. Springer, 2016, pp. 351–363.

[10] T. Matsuo, T. Sonoda, Y. Takemura, M. Sato, and K. Ishii, "Toward smart tomato greenhouse: The fourth tomato harvesting robot competition," *Journal of Robotics, Networking and Artificial Life*, vol. 6, no. 2, pp. 138–142, 2019.

[11] The ROSE challenge. [Online]. Available: http://challenge-rose.fr/en/home/

[12] F. Amigoni, E. Bastianelli, J. Berghofer, A. Bonarini, G. Fontana, N. Hochgeschwender, L. Iocchi, G. Kraetzschmar, P. Lima, M. Matteucci *et al.*, "Competitions for benchmarking: Task and functionality scoring complete performance assessment," *IEEE Robotics & Automation Magazine*, vol. 22, no. 3, pp. 53–61, 2015.

[13] ACRE evaluation plan. [Online]. Available: https://metricsproject.eu/wp-content/uploads/2020/07/Submitted-deliverables_M7_METRICS-deliverable_D5.1-ACRE_evaluation_plan_post-review2.pdf

[14] G. Avrin, V. Barbosa, and A. Delaborde, "Ai evaluation campaigns during robotics competitions: the metrics paradigm," in *1st international workshop on Evaluating Progress in Artificial Intelligence (EPAI 2020) in conjunction with ECAI 2020*, 2020.

[15] O. Galibert and J. Kahn, "The first official repere evaluation," in *First Workshop on Speech, Language and Audio in Multimedia*, 2013.

[16] (2020) Acre dry-run cascade competition. [Online]. Available: https://competitions.codalab.org/competitions/27176

[17] P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.