ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Improving Volume Prediction of Wheat based on multi-view Images

Semester project

Yann Bourdé

November 10, 2025

Advisors: Prof. Dr. Hans-Andrea Loeliger, Olivia Zumsteg, Dr. Paraskevi Nousi, Dr. Lukas Roth, Dr. Norbert Kirchgessner

Department of Electrical Engineering and Information Technology, ETH Zürich

Abstract

Accurate, non-destructive estimation of wheat spike volume is crucial for high-throughput phenotyping of yield-related traits under field conditions. The thesis "Multi-View Deep Learning for 3D Wheat Spike Volume Estimation in the Field" proposed a multi-stage pipeline based on images from the Field Phenotyping Platform (FIP), combining spike detection, multi-view clustering, segmentation, feature extraction and deep learning-based regression. This project extends that work by analysing the detection and segmentation stages and their impact on the final volume prediction. We compare YOLOv5, YOLOv11 and YOLOv12 models trained on GWHD data for spike detection on FIP2 images, evaluating accuracy, inference speed and their effect on downstream clustering and volume estimation. All detectors achieve comparable detection performance; additional recovered views only marginally influence prediction accuracy, supporting the choice of efficient models such as YOLOv11. For segmentation, we train YOLOv11 on cropped auto-masked spikes and assess its influence on LSTM- and transformer-based volume models under different mapping and normalization strategies. The results show subtle but task-dependent effects, highlighting the importance of consistent input scaling and robust segmentation. Overall, our findings provide practical guidelines for designing an accurate, efficient and scalable spike volume estimation pipeline.

Acknowledgments

I would like to thank my direct supervisors Olivia Zumsteg and Dr. Paraskevi Nousi for their help, guidance, constructive feedback, and encouragement throughout this work.

I would like to also thank Prof. Dr. Hans-Andrea Loeliger for taking on the role of academic advisor for this project.

Contents

Chapter 1

Introduction

Assessing the volume of wheat spikes is crucial for evaluating crop resilience, particularly under challenging environmental conditions such as heat and drought. Volume serves as an important indicator of fruiting efficiency, a factor that directly determines overall yield. However, existing phenotyping methods are often slow, labor-intensive, or lack the scalability required for large-scale selection of high-performing wheat varieties. As a result, there is a need for high-throughput approaches that can accurately estimate spike volume in an efficient manner [1].

The thesis "Multi-View Deep Learning for 3D Wheat Spike Volume Estimation in the Field" [1], introduced a system for predicting the volume of individual wheat spikes using images captured by the Field Phenotyping Platform (FIP) [2]. This research project aims to continue the work done in that thesis by exploring its shortcomings and improving the process. In particular, this project focuses on exploring and improving two main areas of the current pipeline which are detection and segmentation. These parts are crucial to the system and their improvement could lead to accuracy and speed improvements for the overal piepline.

1.1 Background

The FIP is a specialized prototype designed for large-scale crop phenotyping. It features a sensor head mounted on a rope-driven carrier system, enabling the simultaneous capture of 13 RGB top-view images of crops [2]. Figure 1.1 illustrates the FIP system along side one of the 13 images of a wheat field taken by the FIP.

Figure 1.1: Visualization of the camera set up and a standard field image [1].

The final process developed in the thesis is presented in Figure 1.2 below and works as follows:

- 12 out of the 13 images taken by the FIP serve as input.
- The input is passed through a deep learning object detection model, locating the wheat spike and outputting a corresponding bounding box.
- A clustering algorithm based on the works of [3] is used to create clusters of spikes in between the 12 views. This means finding the same spike in the 12 different views.
- Each view of the clusters is then passed through a deep learning segmentation model to segment wheat spikes from their bounding box.
- We then run the segmented spikes through DINOv2 [4] and obtain features.
- We then run these features through a transformer based deep learning volume prediction model to get out final volume output in cubic millimeters (mm^3).

At the start of the project, the detection model used was a YOLOv11 medium sized detection model and the segmentation model was a YOLOv11 medium sized segmentation model. Both were pretrained on the COCO dataset [5] and then re-trained on more datasets that are presented in the next chapter.

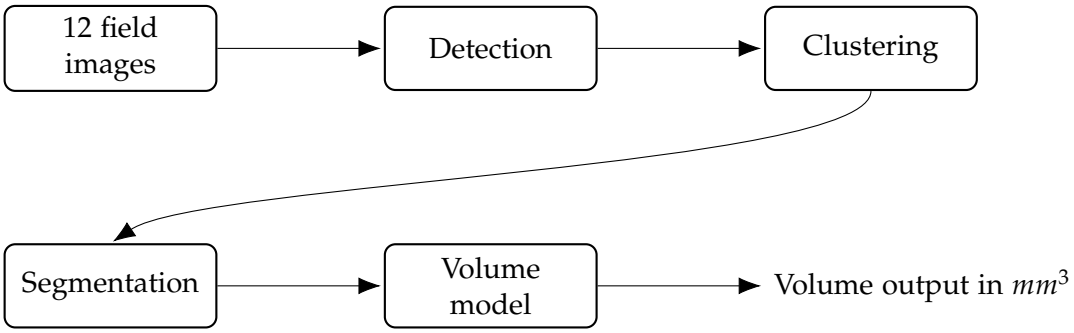


Figure 1.2: Overview of the pipeline from input images to final volume prediction.

1.2 Process' shortcomings

This system has shortcomings that affect different areas of the process. The first area is detection and clustering, the second is segmentation and the third is volume prediction. As mentioned before, in this project we focus on exploring the shortcomings of the first two parts. Those are:

- Detection speed and accuracy.
- Clustering view retention.
- Segmentation speed and accuracy.

In the following chapters the work done on this project is presented in chronological order with discussions on all results and findings.

Chapter 2

Datasets

Throughout the course of this project, three main datasets where used to train and evaluate different models. Two of them, FIP and GWHD were previously used in the thesis and the last one was recently made. All three are presented below.

2.1 FIP

The dataset was collected over two years (2023 and 2024) on six different sampling dates in total. These six dates correspond to three distinct growth stages, as the sampling dates in 2023 and 2024 represent the same growth stages. Around 1700 spikes have been tagged and imaged in the field. After that, the spikes were collected and a 3d-scan was created, using the Shining 3D Einscan-SE V2 [6] 3D light scanner, in order to obtain ground truth volume. In the thesis, tagged spike were manually marked, since it proved difficult to detect the tags in the first place, and also to infer which tag actually belongs to which spike. Roughly 200 images had to be removed manually from the dataset for different reasons, e.g., because it was not possible to find their tag in the images or to clearly associate a single spike with its label. In total, around 700 spikes were allocated for training, 150 for validation, and 250 for testing - even though this number varies based on the experiments. To minimize genotype-specific overfitting, the data split was performed at the genotype level.

2.2 GWHD

The GWHD dataset [7] is a large-scale dataset containing bounding box annotated wheat spikes from different continents. It contains 6512 images of size 1024x1024 with over 300k bounding boxes in total. For our experiments, we used a train/val/test split of 5091/40/1381. This dataset served as the

main training dataset for our detection and segmentation experiments thanks to its labeled bounding boxes.

2.3 FIP2 manually labeled dataset

This dataset features the same characteristics as the FIP dataset, which includes images of plots with 12 different usable views. It includes 7 plots with 36 images each. It also contains one labeled image per plot which results in 7 images with manually labeled bounding boxes for testing. At the time of this project, this was the only FIP dataset manually labeled with bounding boxes, and it was used for the detection experiments.

Chapter 3

Detection Experiments

Through the course of this project, there were 4 main detection experiments performed:

- **Experiment 1:**
  Performance evaluation of different detection models on the FIP2 manually labeled dataset.
- **Experiment 2:**
  Timing analysis of different detection models on the FIP2 manually labeled dataset.
- **Experiment 3:**
  Evaluation of clustering performances with different detection model.
- **Experiment 4:**
  Performance evaluation of different detection models on the final volume output.

3.1 Experiment 1

The first experiment was about evaluating the accuracy performance of different detection models on labeled FIP data, with the motivation of exploring different model sizes. For this experiment, we used three different models:

- YOLOv5 (challenge winner) [8]. This model was used as part of a competition on the GWHD dataset.
- YOLOv11 detection model [9] pretrained on COCO dataset [5].
- YOLOv12 detection model [10] pretrained on COCO dataset [5].

Note that the yolo11_train_val_20ep is the baseline model which was used in the system's pipeline before this project. Table 3.1 below shows the different parameter counts of each model.

Models	# Param
yolo5_all_20ep	87,198,694
yolo5_all_40ep	87,198,694
yolo5_train_20ep	87,198,694
yolo11_all_20ep	25,311,251
yolo11_all_40ep	25,311,251
yolo11_train_val_20ep	20,053,779
yolo12_all_20ep	26,389,875
yolo12_all_40ep	26,389,875
yolo12_train_20ep	26,389,875

Table 3.1: Number of parameters for the evaluated detection models.

All models were trained on the GWHD dataset and then tested on the FIP2 dataset. We tested with different training set sizes by including or excluding the validation and testing sets of GWHD for training the models. We also experimented with different training length by varying the number of epochs. Table 3.2 below highlights the results.

	Macro				Micro					
Models	**Precision**	**Recall**	**F1**	**AP50**	**Precision**	**Recall**	**F1**	**TP**	**FP**	**FN**
yolo5_all_20ep	0.73	0.91	0.81	0.68	0.73	0.91	0.81	1262	458	124
yolo5_all_40ep	0.73	0.90	0.81	0.68	0.73	0.90	0.81	1251	455	135
yolo5_train_20ep	0.67	0.89	0.77	0.62	0.68	0.89	0.77	1235	587	151
yolo11_all_20ep	0.76	0.93	0.84	0.72	0.77	0.93	0.84	1285	393	101
yolo11_all_40ep	**0.77**	**0.94**	**0.84**	**0.74**	**0.77**	**0.94**	**0.84**	**1300**	**393**	**86**
yolo11_train_val_20ep	0.68	0.85	0.76	0.59	0.69	0.85	0.76	1181	544	205
yolo12_all_20ep	0.75	0.91	0.82	0.70	0.75	0.91	0.82	1258	419	128
yolo12_all_40ep	0.75	0.93	0.83	0.71	0.76	0.92	0.83	1280	409	106
yolo12_train_20ep	0.67	0.88	0.76	0.60	0.68	0.88	0.76	1218	583	168

Table 3.2: Detection performance of different detection models variants (macro and micro metrics).

We evaluate the performance using different metrics. Macro metrics are arithmetic means of per-image values and micro metrics are computed from global sums (TP/FP/FN). From these results, the best performing model is the YOLOv11 large model trained on all of the GWHD images for 40 epochs. It is notable however that this performance is closely followed by some other models such as the YOLOv5 trained on all data for 40 epochs and the YOLOv12 trained on all data for 40 epochs as well.

3.2 Experiment 2

In this second experiment, the aim was to compare the models on inference speed as it is also a key aspect of this system's pipeline. Table 3.3 below shows the results.

Models	plot_461	plot_462	plot_463	plot_464	plot_465	plot_466	plot_467	Avg (ms)
yolo5_all_20ep	1127	1121	1131	1143	1123	1133	1106	**1126**
yolo5_all_40ep	1124	1250	1180	1143	1076	1080	1106	**1137**
yolo5_train_20ep	1119	1120	1123	1141	1116	1134	1098	**1122**
yolo11_all_20ep	1348	1270	1348	1264	1263	1277	1260	**1290**
yolo11_all_40ep	1352	1357	1370	1263	1253	1272	1263	**1304**
yolo11_train_val_20ep	1976	1169	1169	1158	1167	1168	1936	**1392**
yolo12_all_20ep	3261	3315	3328	3307	3313	3321	3316	**3309**
yolo12_all_40ep	3313	3315	3316	3316	3321	3327	3316	**3318**
yolo12_train_20ep	3302	3285	3286	3302	3306	3302	3300	**3298**

Table 3.3: Average inference time of different detection model on the FIP2 dataset.

Each value corresponds to the average inference speed on one image. Even if only one image per plot was labeled, the inference was run on all images available in each plot which accounts for 36 images per plot. The global average was then computed and is presented in the last column. We see that the fastest model is YOLOv5 followed by YOLOv11 and YOLOv12 respectively. We see a large gap between yolo11 and yolo12 with almost double the inference speed for the latter.

3.2.1 Discussion - Experiments 1&2

The interesting observation from these two experiments is that the overall accuracy is close for all models but the inference speed is significantly different. YOLOv5 is the fastest which is expectable as its architecture is designed for speed [cit] but is closely followed by YOLOv11. YOLOv12 on the other hand is much slower and does not offer any accuracy advantages in this setting. Additionaly, we observe that there is a large difference in performance between models trained on all the data and the models trained only on the training set. This could indicate that the training data is not large enough at the moment to max out the performance of these models for this task.

In terms of parameters, YOLOv5 contains the most which makes it the less portable model out of all three – an important consideration for deploying the model on edge devices. Combining all these results, it is safe to say the YOLOv11 large model is the best model tested for this task as it combines speed and accuracy while remaining portable with a weight file of 50MB (160MB for YOLOv5 and 50MB for YOLOv12).

3.3 Experiment 3

The next experiment is about evaluating the performance of different detection models in retaining views after the clustering algorithm is used. The clustering algorithm, based on the works of [3], groups bounding boxes

together. Consequently, the performance of the detection model can affect the number of views present in a cluster. Here, views refer to the number of different camera angle which have a bounding box (view) for a given spike. For instance, a spike might be detected in only 10/12 images for a given plot and the clustering algorithm might only group 8 out of those 10 available bounding boxes. Consequently, if a detection model can detect more views of a same spike, it might result in larger clusters if the clustering algorithm can also group them together.

For this experiment, we tested using 3 models: the baseline model (yolo11_train_val_20ep), the best performing model from our first experiment (yolo11_all_40ep) and the best performing YOLOv5 model (yolo5_all_20ep). We did not experiment using YOLOv12 since the model was already performing worse than YOLOv11 and does not have an inference speed advantage as revealed in the first two experiments. The results from this experiment are presented in Table 3.4 below.

Models	gt_views	cluster_views	missing_in_gt	present_in_gt	Avg. vol. diff. (mm^3)
baseline – yolo11_train_val_20ep	1969	1994	79	1915	1340
yolo11_all_40ep	1969	2043	109	1934	1335
yolo5_all_40ep	1969	2020	90	1930	1317

Table 3.4: Clustering performance with different detection models.

The ground truth in these experiments refers to the number of views present in the mapping file of the FIP dataset. This mapping file includes which plants and which images (1 to 12) was ultimately used to train, validate and test the volume prediction models - and it was made manually.

The performance is evaluated using different metrics. There is the total number of views from all clusters, the total number of views not present in the ground truth and the total number of views present in the ground truth. A volume metric is also present, but it is for Experiment 4 in Section 3.4.

We observe that all models are able to find new views that were not present in the ground truth but yolo11_all_40ep is the one that is able to find the most with 109 views missing from ground truth. However, we also observe that all models fail to find all views present in ground truth, with the baseline model being the worst with 1915/1969 views. What we can conclude is that there is a possibility of finding more views using different detection models which means that we can potentially give more information to the volume prediction model and thus obtain a more accurate volume prediction. We test this intuition in the following experiment.

3.4 Experiment 4

In this experiment, we try to see if using the extra views from different detection model results in better performance for volume prediction. To do so, we simply use the clusters found in the previous experiment and feed them through the volume detection model like in the normal pipeline. The results are highlighted in Table 3.4 in Experiment 3.

Here we can make a very interesting observation, the average volume error is almost the same for all three models. This result indicates that extra views do not necessarily translate into a more accurate volume prediction. This is most likely due to the fact that the marginal increase in information from one extra view is small. For instance, if a cluster already contains 8-11 views, one or two extra views might not be enough to significantly improve volume prediction.

3.5 Key Takeaways

From these experiments, we can highlight a few key takeaways:

- Newer detection models do not necessarily result in increased accuracy.
- Additional training data leads to a significant increase in performance for these models in this setting.
- Smaller models provide very good inference speed while retaining accuracy.
- Different detection models do not significantly influence the final volume output, even if more views can be found.

Chapter 4

Segmentation Experiments

The second aspect of this project focuses on the segmentation part of the system's piepline developed in the thesis. Since segmentation provides the final input to the volume model, the intuition is that a better segmentation model will provide more accurate and consistent segmented wheat spikes to the model for volume prediction. Unfortunately, there is no large labeled segmentation dataset for wheat spikes (at the time of this project) which makes it challenging to train a segmentation model.

The solution found in the thesis was to use SAM2.1 [11] to provide rough masks to the GWHD dataset since SAM can work with detection boxes as input. A YOLOv11 medium segmentation model was then trained on this auto-masked GWHD dataset. After experimenting with variations of SAM2.1 and trying to train a YOLOv12 segmentation model from scratch, no conclusive improvement was made so the focus switched to another idea.

The old segmentation model was trained on 1024x1024 auto-masked images, however, in the pipeline, segmentation is done on bounding boxes with single wheat spikes inside them. The idea was then to crop the auto-masked dataset and retrain a YOLOv11 segmentation model and compare their performances with the intuition that a model trained on image data closer to what it would see during inference would lead to more accurate segmentations and therefore more accurate volume predictions. Figure 4.1 below shows an example of qualitative improvements from using the new segmentation model trained on cropped auto-masked GWHD.

Since there are no manually labeled ground truth mask values for this dataset, it is hard to evaluate the performance across multiple images. From manually analyzing roughly 30 images, it was visible that the new segmentation model could create less "patchy" masks and follow the "line" of the

Old model New model

Figure 4.1: Comparison between old and new segmentation models on an image.

wheat spikes better. Once again, this is a visual evaluation and is therefore not very reliable.

Fortunately, the unsegmented dataset made of images corresponding to cropped detection boxes from the FIP data was available to us. Consequently, we could try and segment this dataset using the two segmentation models and then use these new segmented datasets to train volume prediction models (LSTM and regulated transformer) to see if we could obtain an increase in performance. Figure 4.2 below highlights the proceeding for this experiment.

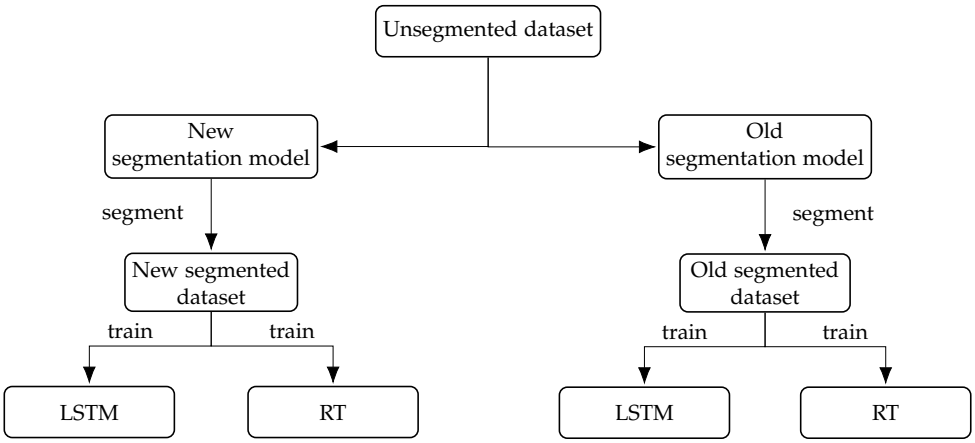


Figure 4.2: Overview of how the new and old segmentation models generate datasets used to train LSTM and RT volume models.

A technicality of this experiment is the common mapping versus own mapping. This refers to the fact that segmentation models can fail to segment resulting in different sets of segmented images depending on which segmentation model is used. Thus, we performed experiments on common mappings (union of sets) and own mappings - where volume models are trained on all the data from their respective segmented datasets. Furthermore, we train and test the models with and without distance normalization to see if it affects results. Distance normalization is used to re-scale the images as if they were taken from equal distances and it has been shown to improve volume prediction in [1].

The models are evaluated using three metrics: Mean absolute error MAE, Correlation ρ and Steepness s. Steepness is the steepness of a linear regression between the predicted and ground truth volume. Equations 4.1, 4.2 and 4.3 show the formal definition of all three metrics, where x_i is the i-th ground truth, y_i the i-th prediction, $\bar{x}$ and $\bar{y}$ are the mean over all x_i resp. y_i and n is the total number of predictions.

$$MAE = \frac{1}{n}\sum_i |x_i - y_i| \tag{4.1}$$

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{4.2}$$

$$s = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} \tag{4.3}$$

In the following figures, new_data represens the dataset segmented using the new segmentation model (trained on cropped auto-masked images) and old_data represents the dataset segmented using the old segmentation model (trained on full auto-masked images). Similarly, new_lstm / old_lstm /new_RT /old_RT indicate with which data each model was trained on. We used the same training process and architectures for the LSTM and RT training as in the thesis. The goal is to have a level playing field to compare the peformance of each segmentation model on volume prediction.

4.1 Common mapping results

4.1.1 Common mapping - LSTMs

No distance normalization

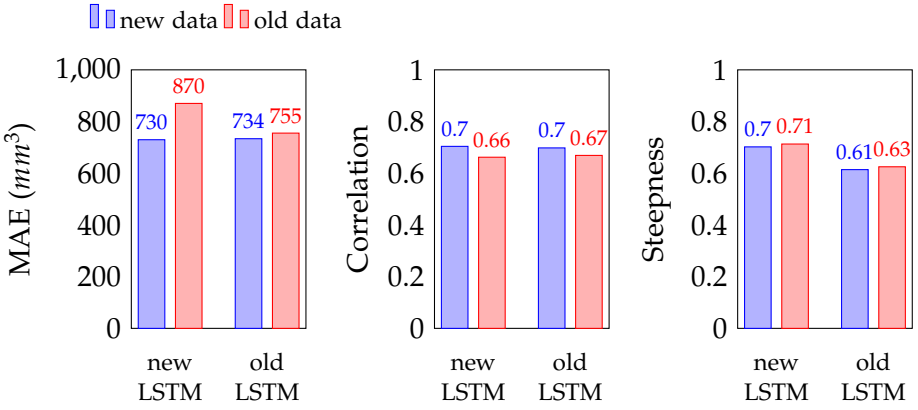


Figure 4.3: LSTM models performance using common mapping without distance normalization.

We observe that MAE performance is better for the old model even though the gap is small on the new data. Correlation values are similar and steepness is slightly better for the new model.

With distance normalization

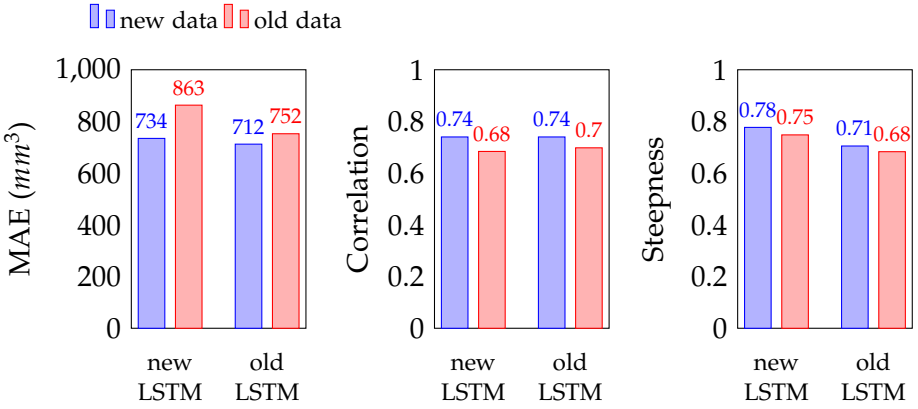


Figure 4.4: LSTM models performance using common mapping with distance normalization.

We observe that the MAE is a little better for the old model. Correlation values are similar but both models get higher values on the new data. Steepness is slightly better for the new model.

4.1.2 Common mapping - Regulated Transformers

No distance normalization

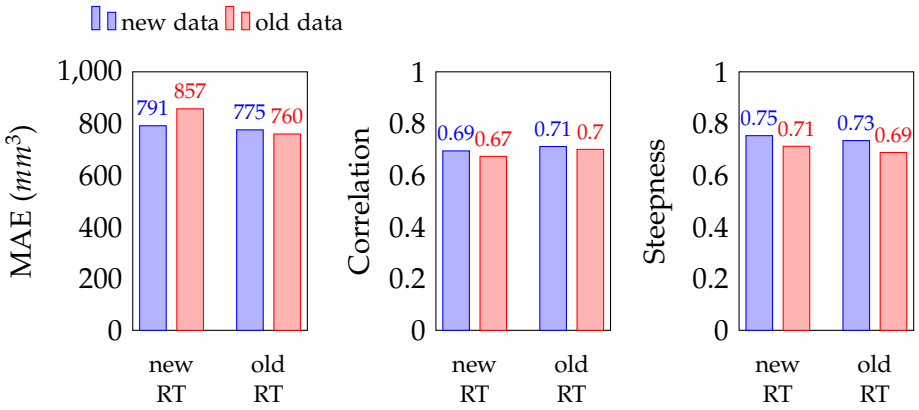


Figure 4.5: Regulated Transformer models performance using common mapping without distance normalization.

We observe that MAE performance is better for the old model. Correlation is also better for the old model but steepness favors the new model.

With distance normalization

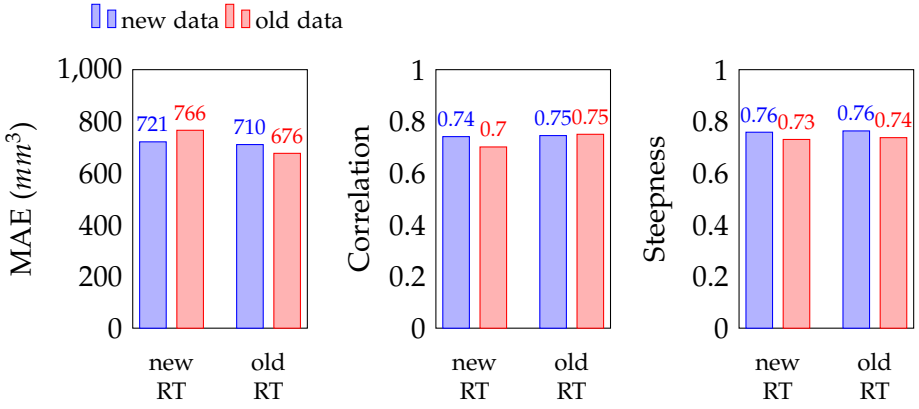


Figure 4.6: Regulated Transformer models performance using common mapping with distance normalization.

We observe slightly better MAE for the old model. Correlation is significantly better for the old model and steepness is about the same for both models.

4.2 Own mapping results

4.2.1 Own mapping - LSTMs

No distance normalization

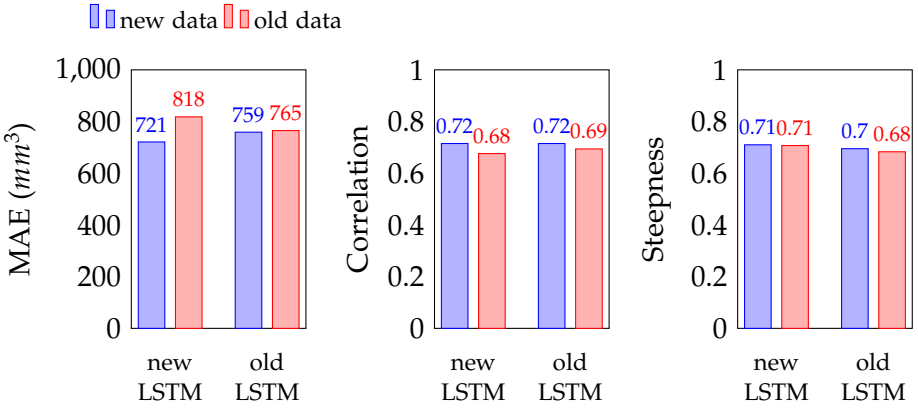


Figure 4.7: LSTM models performance using own mapping without distance normalization.

We observe better MAE performance of the new model on its data but much worse performance on the old data. The old model does consistently good on both data. Correlation is approximately the same on the respective datasets and steepness favors the new model.

With distance normalization

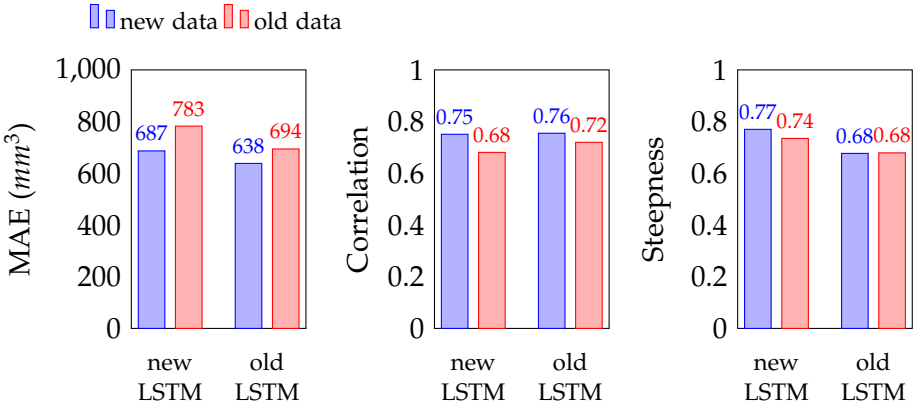


Figure 4.8: LSTM models performance using own mapping with distance normalization.

We observe slightly better MAE on the old model. Correlation is slightly better for the old model and steepness is better on the new model.

4.2.2 Own mapping - Regulated Transformers

No distance normalization

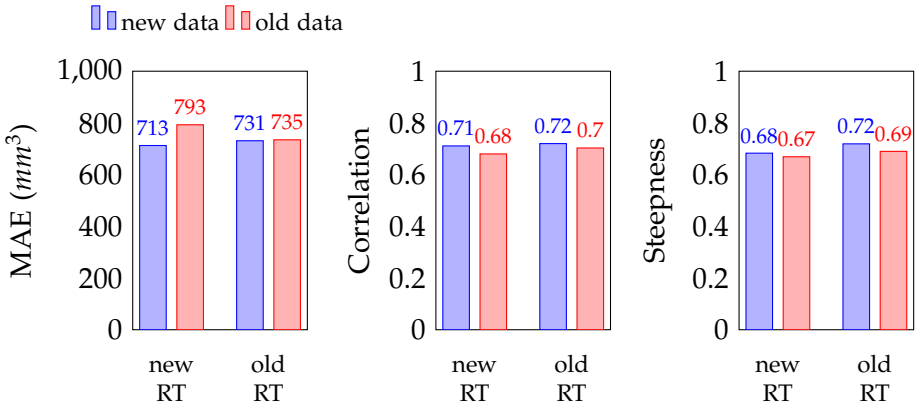


Figure 4.9: Regulated Transformer models performance using own mapping without distance normalization.

We observe better MAE performance for the new model on its data but much worse performance on the old data. The old model does consistently good on both data. Correlation is slightly better for the old model and steepness is also better for the old model.

With distance normalization

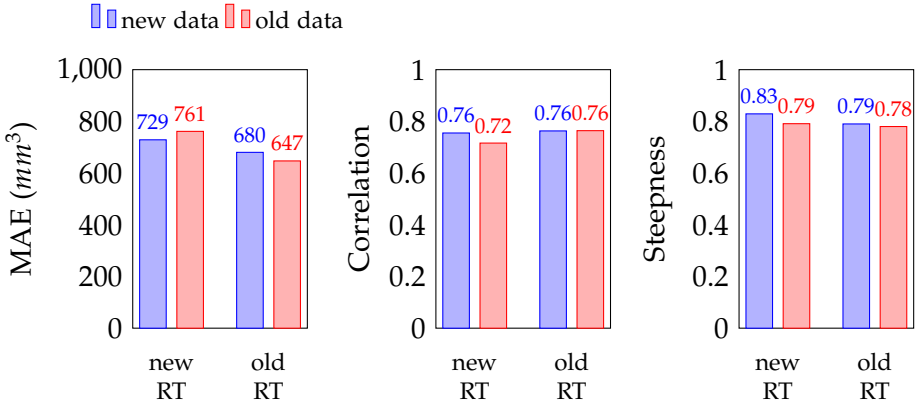


Figure 4.10: Regulated Transformer models performance using own mapping with distance normalization.

We observe better MAE performance on the old model. Correlation is also better for the old model and steepness is better for the new model.

4.3 Result discussion

The results from these experiments are very interesting because they show that the new and old models usually perform similarly on the new data. On the old data however, the new models fail to perform well. This could indicate that the new models 'overfit' to their data and are unable to generalize to the old data. Moreover, if we consider the new models' segmentations to be more 'true', then it means that the old models are able to infer quite well on data that is more accurate than their training data, but the other way around does not seem possible. With regards to distance normalization, it is clear that including it significantly improves performance across the three metrics. This is true for both common mappings and own mappings, with LSTMs and regulated transformers. Consequently, all subsequent experiments were performed with distance normalization only.

Regarding the correlation metric, it tends to be higher on the new data for both model types. This is also true for steepness even thought it is usually a little higher with the new models. This could indicate that the new data (made using the new segmentation model) includes less cases of spikes being 'far' away from their ground truth segmentation (if such ground truth existed). This leads to trends in data being more recognizable i.e. larger and smaller heads are viewed as such - and more distinctively - leading to higher correlation. This is also visible in steepness because a steepness closer to 1 indicates that the models underpredict by a lesser amount on average, which could also be a result of more accurate and consistent segmentation.

However, we also observe that the differences are usually very small, so the improvement is not very consequent. Moreover, the old models tend to have lower absolute MAE's which is important.

4.4 Further Segmentation Discussion

Late into the project, it was discovered that the effective size of the wheat spikes seen by the segmentation model was not the same for the new and old model. This is because the old model was trained on 1024x1024 images with wheat spikes having bounding boxes of approximately 150x150 and a image rescale parameter of 640. This results in an effective size of of $150 \times 640/1024 = 94$x94 for the old model during training. For the new segmentation model the image is the size of the bounding box since we used cropped images. With a training parameter of 384, this results in effective wheat spike size of $150 \times 384/150 = 384$ during training.

During inference, the images are of size 300x300 with black padding surrounding the images which is usually about 55% of the image. The

inference image size parameter was set to 288 (following the pipeline) which gives and effective size of $165 \times 288/300 = 158$px. This means that the old model was given images that appeared zoomed in (94 to 158) and the new model zoomed out (384 to 158).

Consequently, we ran some more experiments using different image size parameter values and compared with the previous results. This was only done using own mapping and with distance normalization on LSTM training and testing. With an image size parameter of 544 (effective $= 165 * 544/300$ 300) which is closer to what the new segmentation was used to see during training, the LSTM trained on this data obtained a MAE of 640 mm^3 on the new data which corresponds to the performance of the old model on the old data which was the best performance recorded for this category of experiment (see Figure 4.8).

Due to time constraints, this area of the project was not explored more but it is definite that effective wheat size can play a larger role in improving segmentation.

4.5 Segmentation Speed Experiment

The last experiment of this project was to compare different segmentation model trained on the same data but with different sizes (parameters) to evaluate their speed versus accuracy. To do this, we took the old segmentation model base (YOLOv11 medium segmentation) and compared it with its nano, small, large and extra large (xlarge) variants. All models were retrained on the same data that the old segmentation model was trained on, and the they were evaluated on the validation set. Table 4.1 below shows the results.

Model	Size (MB)	Latency (ms)	FPS	mAP@50	mAP@50–95	Precision	Recall
XLarge	119	23	43	0.94	0.78	0.94	0.87
Large	53	15	66	0.94	0.77	0.91	0.88
Medium	43	12	85	0.93	0.76	0.94	0.85
Small	20	9	110	0.93	0.73	0.92	0.85
Nano	6	8	122	0.90	0.68	0.91	0.81

Table 4.1: Size, latency, and segmentation performance with different segmentation model sizes.

As expected, the larger models are slower but more accurate and it is the opposite for the smaller models. To push this experiment further, we decided to segment the unsegmented dataset using each segmentation model (same as in the previous volume experiments) and then compare the performance of the LSTMs trained on each dataset. The experiment was only performed on LSTMs with distance normalization and own mapping. The amount of failed segmentation is also displayed this time around (it was not a big issue in the

previous experiment). Finally, the performance of each model is evaluated on its respective dataset. Figure 4.11 below shows the results.

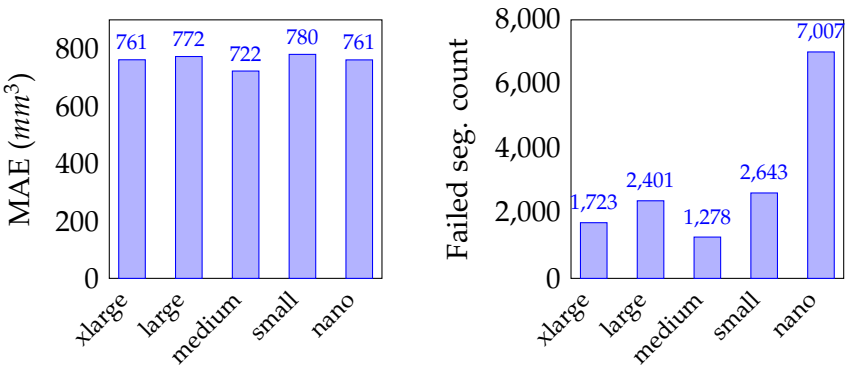


Figure 4.11: LSTM MAE performance and failed segmentation count with different segmentation model sizes.

Interestingly, we see that the MAE is best for the medium sized model and gets worse as the models become smaller and larger. For the smaller, model it could be that they are not large enough to be effective. For the larger models, it is possible that the dataset used in training is not large enough to leverage their size. In any case, appart from the medium, the MAE values are quite close to each other, indicating that the model size does not impact performance that much.

Another interesting result is the number of failed segmentations of each models when making the dataset to train the LSTM. Once again, the medium model performs best but the number of failed segmentation skyrockets for the nano model with around 7000 failed segmentations for 17000 images. The small model also performs twice as bad as the medium. This indicates that the segmentation model must be of a minimum size if it needs to be effective, so choosing a faster but smaller model might be too detrimental in this case.

Chapter 5

Conclusion

The goal of this project was to improve an existing multi-view deep learning pipeline for estimating wheat spike volume from Field phenotyping Platform (FIP) images by focusing on two crucial components: detection and segmentation. Building on the work of [1], we evaluated how different architectural choices and training strategies in these components affect both intermediate performance (detection quality, clustering behaviour, segmentation quality) and the final volume prediction.

From the detection experiments, several clear conclusions emerge. Firstly, when trained on the GWHD dataset, all evaluated models (YOLOv5, YOLOv11, YOLOv12) achieve similarly high detection performance on FIP2. The marginal gains in accuracy between models are small compared to the differences in inference speeds and model sizes.

Secondly, experiments on clustering show that different detection models can recover additional views that are missing from the original ground truth. However, volume prediction model from using these clusters show that additional views do not lead to a meaningful reduction in volume error. Taken together, these results support the choice of a detector that prioritizes inference speed and portability while maintaining solid accuracy, such as the YOLOv11 large model used in this work.

The segmentation experiments show that training a YOLOv11 segmentation model on cropped auto-masked images - which is closer to the single-spike inputs used in the pipeline - led to visually cleaner masks and slightly more consistent trends in some evaluation settings.

By comparing LSTM and regulated transformer models trained on datasets generated with the old and new segmentation models under both common and own mappings, and with and without distance normalization, we observed that:

- The new segmentation can yield comparable or marginally improved correlation and steepness on its own data.
- The old models remain more robust across different datasets.
- Absolute MAE differences between configurations are generally small.

Additionally, the analysis of the effective spike size showed that mismatches between training and inference scales can mask potential gains and that aligning these scales is crucial for this pipeline. Lastly, the segmentation speed experiment demonstrated that neither the smallest nor the largest models are optimal in this setting: the YOLOv11 medium model strikes the best balance between segmentation quality, inference speed and failure rate, whereas overly small models lead to many failed segmentations and overly large models offer no commensurate benefit with the current training dataset.

Overall, the findings of this project lead to three main conclusions:

- Detection is not the primary bottleneck once a reasonably good model is used; it can be chosen to prioritize speed and deployability without sacrificing accuracy.
- Segmentation is more sensitive and directly influences the reliability and consistency of the volume predictions. Improvements require more careful consideration of input scaling, data generation and model capacity.
- The interplay between all components matters: design choices in detection and segmentation should be evaluated in terms of their end-to-end impact on volume prediction, not only via intermediate metrics.

These insights provide practical guidelines for future developments of the multi-view deep learning pipeline for estimating wheat spike volume from Field phenotyping Platform.

Chapter 6

Future work

For future work, there are three areas that I would explore if I continued to work on this topic and that I did not get the chance to investigate due to time constraints:

1. Training the segmentation model on mixed datasets (cropped and full image). It would be interesting to see if a model trained on such data could leverage the 'best of both worlds' and provide better results.
2. Exploring improvements for the volume prediction model. This is the third part of the process which was not explored in this project and is a place where potential improvements can be found. In particular, it would be interesting to see if using DINOv2 is the best option out there or if there are some other approaches that may lead to better results.
3. Finally, the GWFSS dataset [12] might also be interesting to train a segmentation model on. It contains mostly unlabeled data but it might be leverageable for this topic using techniques of self-supervised training.

Bibliography

[1] Jannis Widmer. "Multi-View Deep Learning for 3D Wheat Spike Volume Estimation in the Field". Master's thesis. Zurich, Switzerland: ETH Zurich, Apr. 2025.

[2] N. Kirchgessner et al. "The ETH field phenotyping platform FIP: a cable-suspended multi-sensor system". In: *Functional Plant Biology* 44.1 (2016), pp. 154–168. DOI: 10.1071/FP16165.

[3] Takuma Doi et al. *Descriptor-Free Multi-View Region Matching for Instance-Wise 3D Reconstruction*. 2020. arXiv: 2011.13649 [cs.CV]. URL: https://arxiv.org/abs/2011.13649.

[4] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2024. arXiv: 2304.07193 [cs.CV]. URL: https://arxiv.org/abs/2304.07193.

[5] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2014.

[6] SHINING 3D. *EinScan-SE V2 Desktop 3D Scanner*. Hangzhou, China. Accessed: 2025-03-18. URL: https://www.shining3d.com/professional-solutions/desktop-3d-scanner/einscan-se-v2.

[7] Etienne 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". In: *Plant Phenomics* 2020 (2020), p. 3521852. ISSN: 2643-6515. DOI: https://doi.org/10.34133/2020/3521852. URL: https://www.sciencedirect.com/science/article/pii/S2643651524000359.

[8] Hanlin Yu and Zhongze Chen. *GWC_solution*. https://github.com/ksnxr/GWC_solution. Accessed: 2025-03-10. 2021.

[9] Glenn Jocher and Jing Qiu. *Ultralytics YOLO11*. Version 11.0.0. 2024. URL: https://github.com/ultralytics/ultralytics.

[10] Yunjie Tian, Qixiang Ye, and David Doermann. "YOLOv12: Attention-Centric Real-Time Object Detectors". In: *arXiv preprint arXiv:2502.12524* (2025).

[11] Nikhila Ravi et al. "SAM 2: Segment Anything in Images and Videos". In: *arXiv preprint* (2024).

[12] Zijian Wang et al. "The Global Wheat Full Semantic Organ Segmentation (GWFSS) dataset". In: *Plant Phenomics* 7.3 (2025), p. 100084. DOI: 10.1016/j.plaphe.2025.100084.

Improving Volume Prediction of Wheat based on multi-view Images

Supervisors: Olivia Zumsteg, Dr. Lukas Roth, Dr. Norbert Kirchgessner (Crop Science) and Dr. Paraskevi Nousi (Swiss Data Science Center)

Start Date: July 2025

Keywords

Deep Learning, Computer Vision, Multi-view Images, High-Throughput Phenotyping, Yield Potential, Climate Change

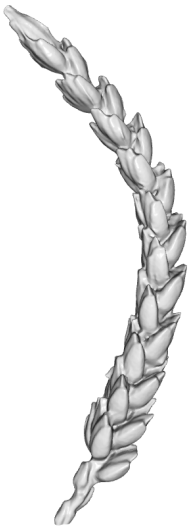
Description

The objective of this thesis is to improve an existing model using deep learning methods to predict the volume of wheat spikes from multi-view images. The spike volume is a promising trait allowing the identification of resilient genotypes under heat and drought (Slafer et al. 2015).

The project will use images of spikes collected with the new sensor head of the Field Phenotyping Platform (FIP) (Kirchgessner et al. 2016) from the Crop Science Lab. This sensor head consists of a rigid setup (https://kp.ethz.ch/infrastructure/FIP.html) of 13 RGB cameras that are triggered simultaneously and capture multiple top-view images from hundreds of different genotypes (Figure 1a). Around 1,800 diverse spikes have been tagged and imaged in the field. The volume of the spikes has been measured in a 3D scanner as ground-truth (Figure 1b) and was be used to train deep learning models.

Models that estimates spike volume based on up to 12 RGB images taken by the FIP (including random viewing angles and changing illumination conditions) have been developed. These models are based on reconstructed 3D point clouds and on original images as input, a pretrained Dinov2 (Oquab et al. 2024), and on an encoder-based transformer.While the models show promising performance, we like to further improve them. We assume that the low quality of the 3D point clouds as input, and missing correspondences of detected wheat heads in images are the major limitation. The aim of this thesis is to improve the prediction accuracy by making architectural or other changes.

(a) New sensor head that captures multi-view images



(b) 3D scan of a wheat spike used to extract the volume

Figure 1: Ground truth data collection

Research Goals

- Improving an existing model trained to predict spike volume based on multi-view images as input.

Expected Outcomes

- Improvement and comparative assessment of the performance and constraints of various deep learning models for the prediction of volume from 2D multi-view images.
- Comparison between the newly developed and existing volume prediction methods.

Impact

Frequent drought and heatwaves due to climate change pose significant threats to global wheat yields. Currently, there is a lack of high-throughput phenotyping (HTP) methods aimed at selecting for fruiting efficiency, i.e., a high number of grains per volume. This thesis aims to enhance the capabilities of ETH's Field Phenotyping Platform (FIP) to accurately predict the volume of wheat spikes, which serves as a key indicator of the yield potential for future wheat genotypes. This trait holds promise in mitigating the impacts of heat and drought under projected climate conditions. Such capabilities are essential to identify resilient genotypes and thereby contribute to global food security.

Find out more about the crop science lab.

Desired Competences

- Python
- Deep Learning and Computer Vision
- Effective collaboration and communication skills
- Interest to experiment and carry out own ideas
- Interest to carry out novel research

References

Kirchgessner, Norbert et al. (Feb. 2016). "The ETH field phenotyping platform FIP: a cable-suspended multi-sensor system". In: *Functional plant biology: FPB* 44.1, pp. 154–168. ISSN: 1445-4416. DOI: 10.1071/FP16165.

Oquab, Maxime et al. (Feb. 2, 2024). *DINOv2: Learning Robust Visual Features without Supervision*. DOI: 10.48550/arXiv.2304.07193. arXiv: 2304.07193[cs]. URL: http://arxiv.org/abs/2304.07193 (visited on 12/19/2024).

Slafer, Gustavo A. et al. (2015). "Fruiting efficiency: an alternative trait to further rise wheat yield". In: *Food and Energy Security* 4.2, pp. 92–109. ISSN: 2048-3694. DOI: 10.1002/fes3.59.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. **In consultation with the supervisor**, one of the following two options must be selected:

☐ I hereby declare that I authored the work in question independently, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].

☑ I hereby declare that I authored the work in question independently. In doing so I only used the authorised aids, which included suggestions from the supervisor regarding language and content and generative artificial intelligence technologies. The use of the latter and the respective source declarations proceeded in consultation with the supervisor.

Title of paper or thesis:

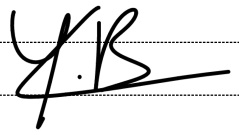
Improving Volume Prediction of Wheat based on multi-view Images

Authored by:
If the work was compiled in a group, the names of all authors are required.

Last name(s):	**First name(s):**
Bourdé	Yann

With my signature I confirm the following:
- I have adhered to the rules set out in the Citation Guidelines.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

Place, date	**Signature(s)**
Zurich, 10/11/2025	

If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.

[1] For further information please consult the ETH Zurich websites, e.g. https://ethz.ch/en/the-eth-zurich/education/ai-in-education.html and https://library.ethz.ch/en/researching-and-publishing/scientific-writing-at-eth-zurich.html (subject to change).