3D Hierarchical Panoptic Segmentation in Real Orchard Environments Across Different Sensors

Matteo Sodano   Federico Magistri   Elias Marks   Fares Hosn   Aibek Zurbayev

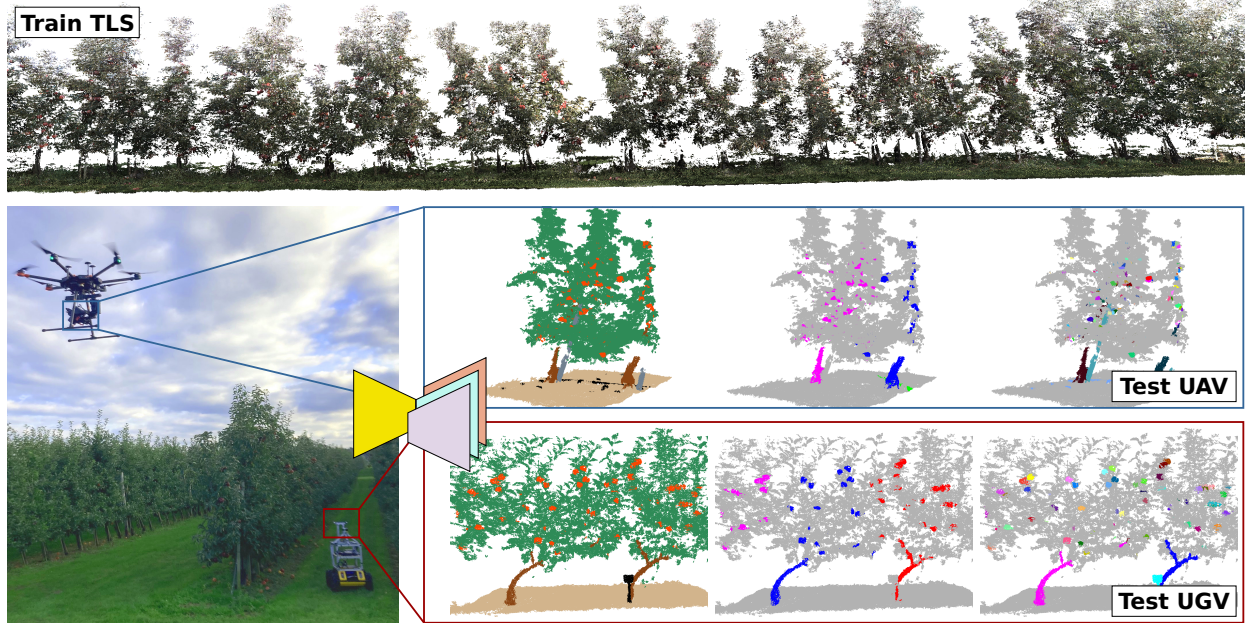Rodrigo Marcuzzi   Meher V. R. Malladi   Jens Behley   Cyrill Stachniss

Fig. 1: View of the apple orchard (top row) recorded with a terrestrial laser scanner (TLS), and used for training our network for hierarchical panoptic segmentation. The test set is composed of different splits corresponding to different sensors and robots. The two examples shown in the image are a UAV equipped with a PhaseOne iXM-100 camera, and a UGV equipped with a RealSense d435i camera.

Abstract— **Crop yield estimation is a relevant problem in agriculture, because an accurate yield estimate can support farmers' decisions on harvesting or precision intervention. Robots can help to automate this process. To do so, they need to be able to perceive the surrounding environment to identify target objects such as trees and plants. In this paper, we introduce a novel approach to address the problem of hierarchical panoptic segmentation of apple orchards on 3D data from different sensors. Our approach is able to simultaneously provide semantic segmentation, instance segmentation of trunks and fruits, and instance segmentation of trees (a trunk with its fruits). This allows us to identify relevant information such as individual plants, fruits, and trunks, and capture the relationship among them, such as precisely estimate the number of fruits associated to each tree in an orchard. To efficiently evaluate our approach for hierarchical panoptic segmentation, we provide a dataset designed specifically for this task. Our dataset is recorded in Bonn, Germany, in a real apple orchard with a variety of sensors, spanning from a terrestrial laser scanner to a RGB-D camera mounted on different robots platforms. The experiments show that our approach surpasses state-of-the-art approaches in 3D panoptic segmentation in the agricultural domain, while also providing full hierarchical panoptic segmentation. Our dataset is publicly available at `https://www.ipb.uni-bonn.de/data/hops/`. The open-source implementation of our approach is available at `https://github.com/PRBonn/hapt3D`.**

I. INTRODUCTION

Crop yield estimation is an important task in agriculture. An accurate yield estimate can help farmers making management decisions to increase crop production and optimize key factors such as harvest time and fertilizer use [16], [41], [44], as well as enable precision interventions [2]. Robots can automatize or support many of these interventions, but to do so reliably and robustly, they need to understand their surroundings through the interpretation of sensor data. Tasks such as segmenting, counting, and localizing fruits in orchards are crucial in horticulture, as they are required for replacing human fruit picking, which is an extremely labor-intensive process [3].

Recently, many approaches targeted perception tasks in agriculture. Semantic segmentation [20], [29] and panoptic

segmentation [26], [34], [43] are especially common. This also motivated the publication of datasets for agricultural perception tasks [12], [13], [42]. However, 3D datasets of real agricultural environments are not common [7], [36].

The main contributions of this paper are twofold. First, we propose an approach based on a convolutional neural network (CNN) for hierarchical panoptic segmentation, which is able to address multiple instance segmentation tasks at once while exploiting their underlying hierarchical relationship by means of a novel skip connections scheme. Second, we introduce a 3D point cloud dataset of real apple orchards, called HOPS (hierarchical orchard panoptic segmentation) recorded with various sensors, such as terrestrial laser scanner, RGB-D cameras, or RGB cameras with subsequent bundle adjustment. Additionally, we release high-quality annotations for hierarchical panoptic segmentation. We recorded data of the apple orchard at different growth stages across several crop rows in the span of two years, and labeled the resulting point clouds in order to obtain semantic segmentation, tree instance segmentation, as well as fruit and trunk instance segmentation. Specifically, in the tree instance segmentation each trunk has a label that is shared with all fruits belonging to it. Thus, in this instance-level task, fruits are not individually labeled. The fruit and trunk instance segmentation yields a different instance for each individual trunk and fruit. We show a portion of the dataset and an exemplary image of the approach we propose in Fig. 1.

Our experiments suggest that (i) our approach can jointly perform semantic, tree instance, and standard instance segmentation on real-world 3D data acquired with different sensors; and (ii) our skip connection scheme allows the CNN to exploit the underlying hierarchical relationship between the individual segmentation tasks, leading to better segmentation performance. Our claims are backed up by the paper and our experimental evaluation.

II. RELATED WORK

Semantic scene interpretation is often required for realizing robotic automation in agriculture, since it identifies task-relevant objects in the scene that enables applications such as monitoring [19], [32], [38] and intervention [1], [15], [22].

Panoptic segmentation [14] unifies semantic and instance segmentation provides semantics of background classes, so-called "stuff", and object-level instances of so-called "things" at the same time. While often images are used for panoptic segmentation in the agricultural domain [43], 3D point clouds generated by RGB-D cameras, LiDARs, or photogrammetric structure-from-motion approaches received increasing research interest [46] due to the capability to extract precise geometric information required for high-precision fruit grasping [15], [22], plant segmentation [30], and fine-grained trait estimation [26]. 3D panoptic segmentation has been actively investigated in the domain of autonomous driving [28], [37], [45], where most approaches use dedicated branches for semantic and instance segmentation using features of a shared encoder. For plant phenotyping, we can exploit the hierarchical structure of plants that can be

decomposed in individual parts [11], such as plant, leaf and even more fine-grained leaf structure. This hierarchical structure has been exploited in the literature, leveraging skip connections between decoders for hierarchical bottom-up instance prediction tasks using images [34], which is the foundation of our approach to 3D panoptic segmentation.

In contrast to prior approaches for 3D panoptic segmentation, we also explicitly consider the hierarchical structure of the prediction task by means of our novel skip connections scheme, which allows us to predict a hierarchical semantic scene interpretation consisting of plant instances as well as fruit instances that can be used for plant monitoring and yield estimation in the agricultural domain. While our approach is applied to orchard scenes, the hierarchical panoptic segmentation framework is general and can be applied to other domains, such as body part segmentation [17].

While most agricultural datasets for semantic interpretation provide only RGB images [21], recently 3D datasets became available. In particular, BUP20 [38] provides an RGB-D dataset for instance segmentation of sweet peppers in a glasshouse [39]. For 3D plant phenotyping in real agricultural fields, BonnBeetClouds [25] provides annotated structure-from-motion point clouds consisting of plant and leaf instance of sugar beet plants. Other datasets provide organ-level annotations of plants using point clouds acquired with high-precision LiDAR scanners [36] or x-ray imaging [7] in the lab. The Crops3D dataset [46] provides organ-level annotations of different crop varieties for point cloud data acquired with a terrestrial laser scanner (TLS) in the field and single plant point clouds using structure-from-motion of RGB images or using structured light cameras.

In contrast to existing agricultural datasets for semantic scene interpretation, we provide a domain-specific agricultural dataset of orchards recorded with different sensors that is annotated with plant and fruit instances. We provide high-resolution point clouds recorded with a TLS, but also structure-from-motion point cloud data acquired with robotic platforms equipped with a high-resolution RGB camera or a consumer-grade RGB-D sensor commonly used in robotics.

III. OUR APPROACH
TO HIERARCHICAL PANOPTIC SEGMENTATION

We propose an approach for hierarchical panoptic segmentation, i.e., the task of simultaneously performing semantic segmentation and multiple instance segmentation tasks with underlying hierarchical relationships. The network we use is an encoder-decoder architecture, and the decoders address semantic segmentation and two-levels of instance segmentation, as illustrated in Fig. 2. The first instance segmentation aims to identify "tree" instances, where a tree is defined by a trunk and all fruits belonging to it. The second instance segmentation looks for all individual instances (i.e., both trunks and fruits) in the point cloud.

A. Architecture

We use a MinkUNet-based neural network [6] for our approach. This kind of neural networks are inspired by the
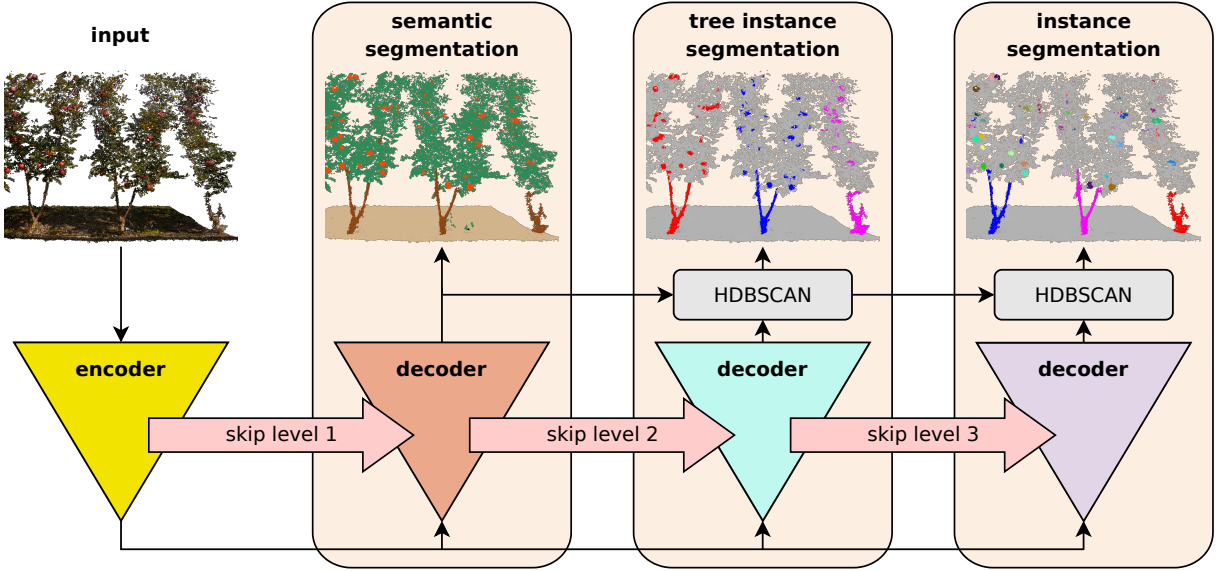
Fig. 2: The encoder takes the colored point cloud as input, and the resulting features are processed by the decoders. We use hierarchical skip connections after each downsampling and upsampling block of the encoder and decoders. We use HDBSCAN to obtain instances.

UNet [35] design, where encoder and decoder are connected by skip connections. To process 3D data, MinkUnet networks use sparse 3D convolutions to process the input data. Specifically, we use the MinkUNet14A model, that allows us to have a lightweight network. In fact, MinkUNet14A is a modification of the standard MinkUNet14 model, and has fewer feature channels per layer, allowing faster computation. We keep the original structure of the network, and replicate the original decoder two additional times. This allows us to have three identical decoders addressing the three segmentation tasks we aim to tackle, which is a convenient solution for exploiting the underlying hierarchy among them.

The first decoder targets *semantic segmentation*, and has a single output head with depth equal to the number of semantic classes that appear in the dataset. It is optimized using the standard weighted cross-entropy loss:

$$\mathcal{L}_{\text{sem}} = -\frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{p} \in \mathcal{S}} \omega_k \, \boldsymbol{t}_p^\top \, \log(\sigma(\boldsymbol{f}_p)), \qquad (1)$$

where $|\mathcal{S}|$ is the number of points in the point cloud, $\boldsymbol{p}$ indicates the individual point, ω_k is a class-wise weight computed via the inverse frequency of each class in the dataset, where $k \in \{1, \dots, K\}$ indicates the semantic class, $\boldsymbol{t}_p \in \mathbb{R}^K$ is the one-hot encoded ground truth annotation at point location $\boldsymbol{p}$, $\sigma(\cdot)$ denotes the softmax operation, and $\boldsymbol{f}_p$ denotes the pre-softmax feature predicted at point $\boldsymbol{p}$.

The second decoder targets *tree instance segmentation*, i.e., the task of finding tree instances where a tree is defined as a trunk and all apples belonging to it. This decoder also has a single output head for offset prediction. Thus, each point predicts a 3D displacement towards a location in the 3D space that facilitates instance separation via clustering. We predict an offset $\boldsymbol{o}_p \in \mathbb{R}^3$ for each point $\boldsymbol{p}$, so that $\boldsymbol{e}_p = \boldsymbol{p} + \boldsymbol{o}_p$ is the 3D location where the point is displaced. To obtain 3D points belonging to the same instance displaced to the same 3D location in space, and points belonging to

different instances displaced to different locations, we adapt the Lovász Hinge loss [31] to 3D as:

$$\mathcal{L}_{\text{tree}} = \frac{1}{|\mathcal{C}|} \sum_{j=1}^{|\mathcal{C}|} \text{Lovász}(\boldsymbol{F}_{\mathcal{C}_j}, \boldsymbol{G}_{\mathcal{C}_j}), \qquad (2)$$

where $\mathcal{C}$ is the set of object instances, $\boldsymbol{G}_{\mathcal{C}_j} \in \{0, 1\}^{|\mathcal{S}|}$ denotes the binary ground truth mask of the j-th instance, and $\boldsymbol{F}_{\mathcal{C}_j}$ is a soft-mask obtained by the offset prediction. The soft-mask $\boldsymbol{F}_{\mathcal{C}_j}$ for instance $\mathcal{C}_j$ is obtained from the offset prediction: each point in the point cloud gets a score that depends on how far from the instance centroid $\boldsymbol{c}_j$ its offset points to. The score is formalized as

$$f_{\mathcal{C}_j} = \exp\left(-\frac{||\boldsymbol{e}_p - \boldsymbol{c}_j||^2}{2\eta^2} \right), \qquad (3)$$

where $\boldsymbol{e}_p$ indicates the 3D location pointed by the predicted offset at point $\boldsymbol{p}$, and η is a hyperparameter that defines an isotropic clustering region around the centroid.

Finally, the third decoder targets standard *instance segmentation*, where we aim to individually segment each trunk and fruit in the scene. Similarly to the previous decoder, here also we have a single output head for 3D offset prediction, and optimize with the Lovász Hinge loss $\mathcal{L}_{\text{ins}}$, as in Eq. (2).

Thus, the final loss function $\mathcal{L}$ is given by a weighted sum of the individual losses:

$$\mathcal{L} = w_1 \, \mathcal{L}_{\text{sem}} + w_2 \, \mathcal{L}_{\text{tree}} + w_3 \, \mathcal{L}_{\text{ins}}. \qquad (4)$$

B. Skip Connections

Skip connections are crucial to ensure feature reusability and address the gradient degradation problem of CNNs. They skip one or more layers, providing a direct gradient flow from the late stages of the decoder to the early stages of the encoder, which is usually hindered by the downsampling operation of the encoder. We aim to extend our previous work on hierarchical panoptic segmentation on RGB images [34]
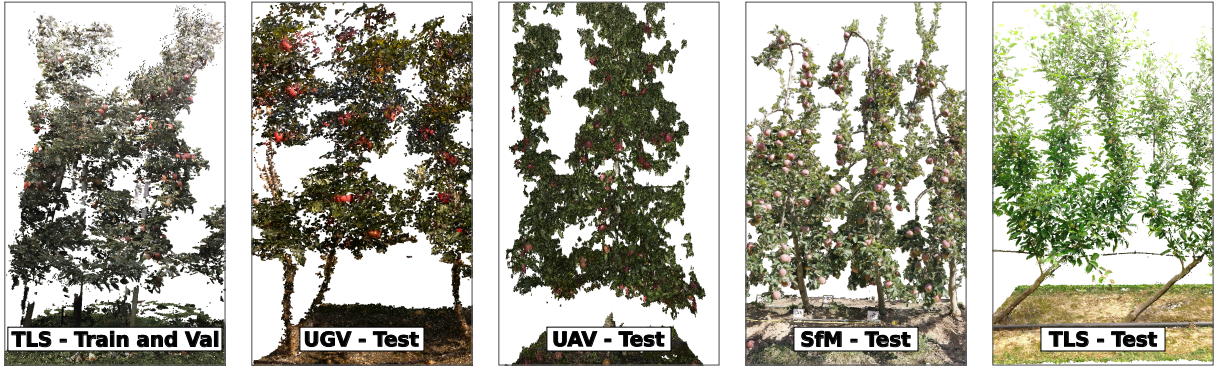
Fig. 3: An example of the colored point clouds of HOPS. The point clouds from the train and validation sets are obtained using a TLS. We have four test sets recorded with different sensors. The one called "SfM" contains point clouds from the Fuji-SfM dataset [9], [10].

to the 3D point cloud domain, and thus we propose to adopt a hierarchical skip connections scheme to address the underlying hierarchy of segmentation tasks. We propose to connect different decoders directly as shown in Fig. 2, instead of encoder and decoders only.

Semantic segmentation. For semantic segmentation, we keep the skip connection from the encoder to the decoder. The spatial information contained in the high-resolution feature maps of the encoder help the decoder for segmentation.

Tree instance segmentation. For tree instance segmentation, we fuse the high-resolution maps coming from the encoder with the feature coming from the same level of the semantic decoder. In this way, the tree instance segmentation decoder will process an enriched feature map, containing also information about semantic segmentation.

Instance segmentation. For instance segmentation, we fuse the high-resolution maps coming from the encoder with the feature coming from the same level of the tree instance segmentation decoder. Notice that, in this way, the skip features include semantic segmentation features as well, that come from the tree instance segmentation branch.

To better illustrate the information flow in our hierarchical panoptic segmentation architecture, consider the process as assembling segmentation in layers. The encoder provides a basic high-resolution spatial guide. The semantic decoder adds labels to it. The tree instance segmentation decoder builds upon this enriched map, identifying individual trees as landmarks. Finally, the instance segmentation decoder refines this by adding finer distinctions, such as fruits and trunks, using all previously gathered information. At each stage, skip connections act like reference notes, ensuring that each level benefits from both low-level detail and high-level context, progressively enriching the representation. This skip connection scheme allows us to exploit the hierarchy among tasks, enriching the features propagated to each decoder. In Sec. V-C, we provide extensive experiments on different skip connection schemes to show that this design choice yields superior performance.

C. Post-Processing

As mentioned, the instance segmentation decoders predict an offset vector for each point. The goal is to have offsets of points belonging to the same instance indicating the same 3D location in space, and offsets of points belonging to different instances indicating different 3D locations. To obtain the final instance masks, we cluster the offsets with HDBSCAN [27]. Additionally, we use the semantic prediction to enforce consistency among instances in the third decoder, for example to avoid the case in which two points with two different semantic predictions end up in the same instance. This cannot be applied to the second decoder, as there instances are composed of one trunk and all apples attached to it, so a single instance actually includes multiple semantic classes.

IV. OUR DATASET FOR HIERARCHICAL PANOPTIC SEGMENTATION

Our dataset, called HOPS, is composed of point clouds collected with three different sensors at Campus Klein-Altendorf near Bonn, Germany. First, we collected point clouds using a TLS placed at multiple locations in the orchard row. Our training and validation set entirely contains point clouds collected with this sensor. Furthermore, we use a few TLS point clouds in the test set, collected in a different year and orchard location compared to the training and validation set. Second, we use point clouds obtained with a bundle adjustment procedure [40] using as input images collected with a UAV equipped with a PhaseOne iXM-100 camera. We flew three missions on the same orchard with a camera angle of 45°, 90°, and 135° from the ground plane at a height of approximately 20 m. This setup allows us to obtain good coverage of the trees including the lower part of the canopy. The photogrammetric point clouds obtained in this way from the UAV are only included in the test set. Third, we use a RealSense d435i mounted on a mobile robot driving through the orchard rows to collect another part of our test set. Again, to obtain photogrammetric point clouds we use a bundle adjustment pipeline. Furthermore, we label a few point clouds from the Fuji-SfM dataset [9], [10] to obtain a fourth test set from a different setup. This dataset is collected in Agramunt in Catalunia, Spain with a Canon EOS 60D DSLR camera, followed by a bundle adjustment procedure to generate point clouds. In contrast to existing datasets in the agricultural domain [12] [33] [42], we specifically designed HOPS to have a remarkable domain shift between train and test sets, to more closely resemble real-world deployments. For this reason, we only included

high-quality point clouds recorded with the TLS in the training set, while the test set includes a variety of sensors.

We report statistics about the different splits of our dataset in Tab. I and we show exemplary point clouds in Fig. 3 to visualize the difference between training and test point clouds. To summarize, our dataset consists of a training set and a validation set obtained with a TLS and four different test set obtained with different sensors (TLS, PhaseOne iXM-100, RealSense d435i, and EOS 60D) in different locations.

For our dataset, we define 5 semantic classes, namely *ground*, *trunk*, *canopy*, *apple*, and *pole*. To formalize the panoptic segmentation task, we define the set of "things" classes (*trunk*, *apple*) and "stuff" classes (*ground*, *canopy*, *pole*). Additionally, we define the *tree* thing class which puts together apples and trunk belonging to the same tree.

To label the data, we split the aggregated point cloud of each orchard row in individual tiles with an average of 1 million points each. We manually label each tile in three stages using the online tool provided by segments.ai. One annotator labels each tile for semantic and two-level instance segmentation, ensuring consistency between instance levels. A second annotator then verifies label quality. Unlike image annotation, point cloud labeling requires frequent viewpoint changes, making it more time-consuming. The average effort consists of 4 h per tile, totaling 525 h for labeling and 175 h for verification.

V. EXPERIMENTAL EVALUATION

We present our experiments to show the capabilities of our method for hierarchical panoptic segmentation on real-world 3D point clouds. The results of our experiments also support our key claims, which are: (i) our approach can jointly perform semantic, tree instance, and standard instance segmentation on real-world 3D data acquired with different sensors; and (ii) our skip connection scheme allows the CNN to exploit the underlying hierarchical relationship between the individual segmentation tasks, leading to better segmentation performance.

A. Experimental Setup

Metrics. For semantic segmentation, we compute the intersection over union (IoU) [8] over all five classes of our dataset, which we discussed in Sec. IV, and report the mean IoU (mIoU) in the following tables. For the tree instance segmentation, we compute the single-class panoptic quality over the "tree" class, that we report as PQ_T. For the standard instance segmentation, we compute the panoptic quality over all classes, and report the mean panoptic quality as PQ in the tables. We also report the overall mean panoptic quality (mPQ) as the average between PQ and PQ_T.

Training details and parameters. In all experiments, we use AdamW [18] with weight decay of 0.99 with an initial learning rate of $5 \cdot 10^{-3}$. We set the weights of the loss function to $w_1 = w_2 = w_3 = 1$. We train our approach for 500 epochs. We set batch size of 1. We use voxel downsampling to $3\,\mathrm{mm}$. Due to practical constraints, the training set

TABLE I: Dataset statistics: number of samples per split, average number of points, fruits, and trunks per sample, and fruits per tree.

	Train	**Val**	**Test**			
	TLS	TLS	UGV	UAV	SfM	TLS
Samples	90	18	12	22	6	27
Points	0.8M	1M	0.6M	1.5M	1.8M	1M
Fruits	90.2	99.2	71.8	99.1	232.3	39.9
Trunks	3.1	3.2	2.9	1.9	2.7	2.9
Fruits/tree	19.2	19.3	17.8	31.7	62.1	9.8

is limited to TLS-acquired data. However, to mitigate potential overfitting to TLS-specific characteristics and improve generalization to non-TLS domains, we apply extensive data augmentation techniques including scale, rotation around all axes, shear, color jittering, and elastic deformation [4]. We tuned all hyperparameters on the validation set.

All baselines can only address one instance segmentation task at a time and, thus, need to be trained twice to solve both instance-level tasks. Our method addresses both instance segmentation tasks with a single training run.

B. Experiments on Hierarchical Panoptic Segmentation

While our PQ is inferior to ForestPS, our approach also yields good performance on tree instance segmentation, on which the dedicated training of ForestPS fails completely, achieving results below 3% on four out of five splits. The first experiment evaluates the performance of our approach and its outcomes support the claim that our approach can jointly perform semantic, tree instance, and standard instance segmentation on real-world 3D data. For this experiment, we compare our approach to existing baselines for panoptic segmentation: MaskPLS [24] is a transformer-based approach that extends Mask2Former [5] to 3D point clouds, while ForestPS [23] is a convolutional neural network approach based on MinkUNet [6], originally designed for tree segmentation in a forest environment. The baselines are trained for only one of the two instance segmentation tasks at a time, as they do not support hierarchical multi-level instance segmentation. As shown in Tab. II for the test set, and Tab. III for the validation set, our approach outperforms the baselines on both splits in terms of mIoU and on most PQ_T, while ForestPS achieves the best PQ on most splits. However, our approach performs better than the baselines on mPQ.

We show qualitative results in Fig. 4. As explained in Sec. IV, only the validation set is recorded in the same time period and with the same sensor as the training set. The test sets, in contrast, are recorded either 2 years later also using a TLS, or with different sensors (UGV and UAV), or belong to entirely different datasets (SfM). This motivates the performance gap between the validation and the test set. We believe that this aspect pushes researchers to build models that yield good generalization performance and can be used in different conditions. In Tab. II, we also report average scores for mIoU and PQ across all four splits and both instance segmentation tasks, to show how different approaches deal with both tasks at the same time.

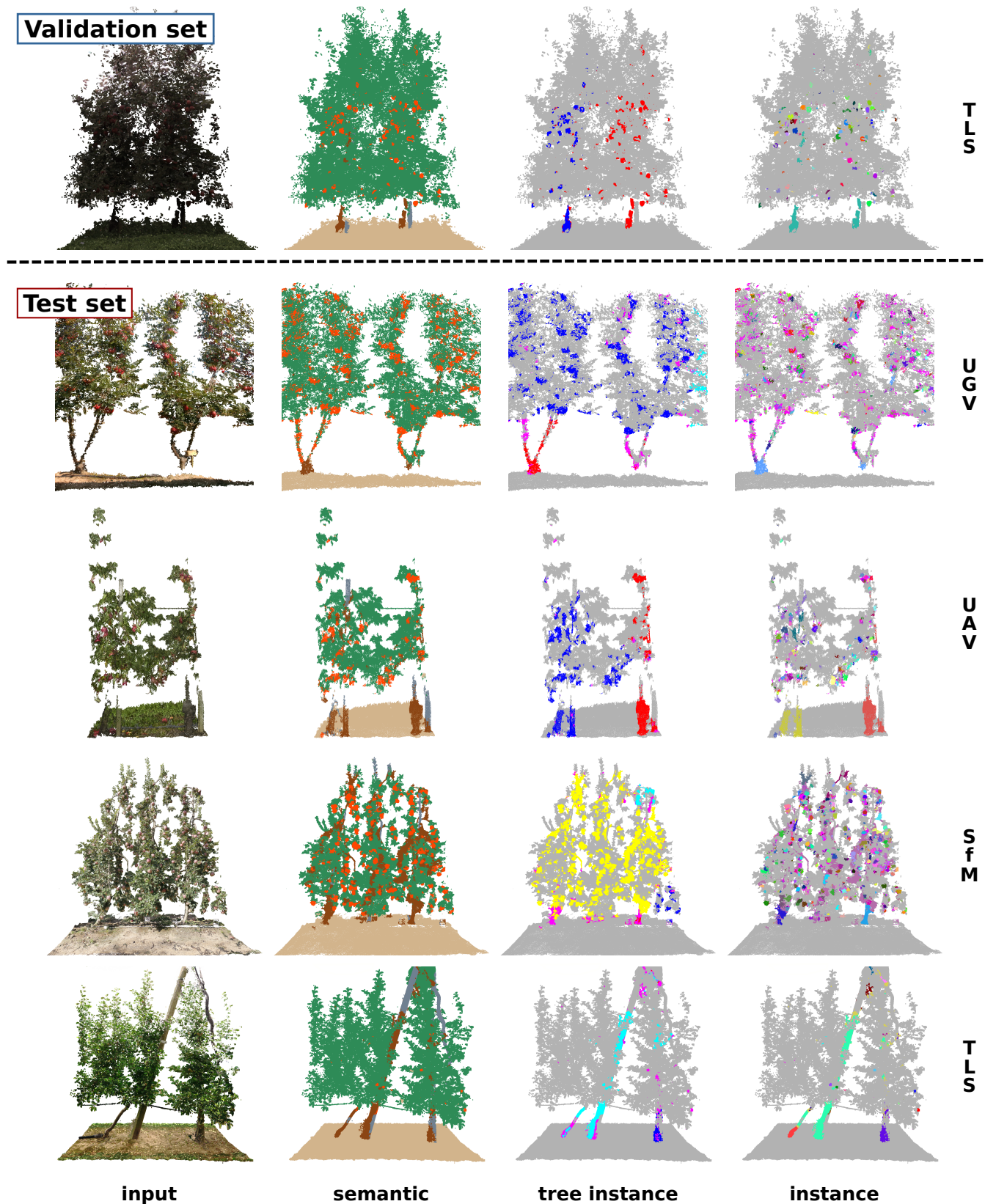Fig. 4: Qualitative results of our approach for hierarchical panoptic segmentation. The first row shows validation set results, the other rows show test set result. We report the input point cloud with all three segmentation predictions, and one result for each sensor. In the semantic prediction, different colors indicate different categories. In the instance predictions, different colors indicate different instances.

TABLE II: Results on the four different test sets of HOPS, and the average score across all sets and all tasks.

Approach	TLS			UAV			UGV			SfM			Average	
	mIoU	PQ	PQ$_T$	mIoU	PQ	PQ$_T$	mIoU	PQ	PQ$_T$	mIoU	PQ	PQ$_T$	mIoU	mPQ
MaskPLS [24]	23.3	36.0	-	20.0	36.0	-	18.7	35.1	-	14.7	41.4	-	23.3	42.3
MaskPLS [24]	-	-	48.9	-	-	48.4	-	-	47.1	-	-	45.3		
ForestPS [23]	52.0	49.4	-	64.3	62.8	-	40.2	37.7	-	47.0	46.2	-	50.9	27.5
ForestPS [23]	-	-	21.5	-	-	2.1	-	-	0	-	-	0		
Ours	57.7	49.5	55.4	65.2	51.9	48.9	42.6	35.2	42.1	47.7	34.5	39.1	53.3	44.6

TABLE III: Results on the validation set. We report the average iteration time (s/it) on a GPU NVIDIA RTX A5000.

Approach	TLS				Param	s/it
	mIoU	PQ	PQ$_T$	mPQ		
MaskPLS [24]	28.2	42.8	-	45.6	1.8M	4.0
MaskPLS [24]	-	-	48.9			
ForestPS [23]	58.9	62.2	-	32.3	1.5M	1.8
ForestPS [23]	-	-	2.3			
Ours	70.9	60.8	52.1	56.5	19.4M	2.0

TABLE IV: Ablation study on the validation set. Comparison between different skip connections schemes.

	Skip Connections	mIoU	PQ	PQ$_T$	mPQ
A	None	55.8	42.8	47.7	45.3
B	Decoder	64.4	53.4	51.3	52.4
C	Encoder	69.8	58.8	50.9	54.9
D	Encoder+Decoder	70.9	60.8	52.1	56.5

C. Ablation Studies

In the ablation study, we aim to validate our claim that the proposed skip connection scheme allows the CNN to exploit the underlying hierarchical relationship between the segmentation tasks, leading to better segmentation performance. We perform this experiment on the validation set. We compared to other skip connection schemes, and report results in Tab. IV. The simplest approach does not use skip connections (denoted as A in Tab. IV), which leads to poor performance. This is expected, since the decoders do not get any high-level information from features coming from earlier stages, which harms gradient flow. The second approach uses skip connections only from decoder to decoder, excluding the contributions from the encoder (denoted as B in Tab. IV). This approach, despite exploiting the hierarchy among tasks with decoder-based skip connections, does not yield good performance. The third approach is the standard UNet-like skip connections from the encoder to all decoders (denoted as C in Tab. IV). In this case, all decoders obtain the same information from the encoder. This proves more effective than the decoder-only skip connections. This is probably due to the fact that restoring high-level information is important, and using decoder-only skip connections does not help gradient propagation. Thus, our method using skip connections from the encoder and from decoder to decoder (denoted as D in Tab. IV) performs better than all others, as on one hand it benefits from the high-level features coming from the encoder, and on the other hand it exploits the underlying hierarchy among segmentation tasks. Interestingly, it also provides better results on semantic segmentation, despite there is no difference in the skip connection scheme for the semantic decoder, as the hierarchy affects only the instance decoders. This suggests that having information coming from the instance decoders indirectly affecting the encoder and the semantic decoder via backpropagation is useful for semantic segmentation performance.

VI. CONCLUSION

In this paper, we presented a novel approach for hierarchical 3D panoptic segmentation on point cloud data. By means of a novel skip connections scheme, our method exploits the underlying hierarchy among different segmentation tasks, and is able to yield semantic segmentation, tree instance segmentation, where a tree is defined as a trunk and all the apples belonging to it, and standard instance segmentation at the same time. Thanks to the proposed skip connection scheme in our architecture, our approach achieves state-of-the-art results, surpassing or being comparable to existing task-specific baselines, despite that they can deal with only one instance segmentation task at a time. Additionally, we introduce a novel point cloud dataset of real apple orchards, called HOPS, labeled for hierarchical panoptic segmentation. Our dataset includes data recorded with different sensors over the course of two years. In our experimental evaluation, we supported all claims made in this paper and showed that our proposed dataset is challenging.

REFERENCES

[1] B. Arad, J. Balendonck, R. Barth, O. Ben-Shahar, Y. Edan, T. Hellström, J. Hemming, P. Kurtser, O. Ringdahl, T. Tielen, and B. Van Tujil. Development of a sweet pepper harvesting robot. *Journal of Field Robotics (JFR)*, 37(6):1027–1039, 2020.

[2] P.M. Blok, F. Magistri, C. Stachniss, H. Wang, J. Burridge, and W. Guo. High-throughput 3d shape completion of potato tubers on a harvester. *Computers and Electronics in Agriculture*, 228:109673, 2025.

[3] L. Calvin. *The US produce industry and labor: Facing the future in a global economy.* DIANE Publishing, 2010.

[4] E. Castro, J.S. Cardoso, and J.C. Pereira. Elastic deformations for data augmentation in breast cancer mass detection. In *Proc. of the Intl. Conf. on Biomedical & Health Informatics (BHI)*, 2018.

[5] B. Cheng, I. Misra, A.G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] C. Choy, J. Gwak, and S. Savarese. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[7] H. Dutagaci, P. Rasti, G. Galopin, and D. Rousseau. ROSE-X: an annotated data set for evaluation of 3D plant organ segmentation methods. *Plant methods*, 16:1–14, 2020.

[8] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *Intl. Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010.

[9] J. Gené-Mola, R. Sanz-Cortiella, J.R. Rosell-Polo, J.R. Morros, J. Ruiz-Hidalgo, V. Vilaplana, and E. Gregorio. Fruit detection and 3d location using instance segmentation neural networks and structure-from-motion photogrammetry. *Computers and Electronics in Agriculture*, 169:105165, 2020.

[10] J. Gené-Mola, R. Sanz-Cortiella, J.R. Rosell-Polo, J.R. Morros, J. Ruiz-Hidalgo, V. Vilaplana, and E. Gregorio. Fuji-sfm dataset: A collection of annotated images and point clouds for fuji apple detection and location using structure-from-motion photogrammetry. *Data in Brief*, 30:105591, 2020.

[11] R. Güldenring, R.E. Andersen, and L. Nalpantidis. Zoom in on the Plant: Fine-Grained Analysis of Leaf, Stem, and Vein Instances. *IEEE Robotics and Automation Letters (RA-L)*, 9(2):1588–1595, 2024.

[12] N. Häni, P. Roy, and V. Isler. Minneapple: a benchmark dataset for apple detection and segmentation. *IEEE Robotics and Automation Letters (RA-L)*, 5(2):852–858, 2020.

[13] J. Kierdorf, L.V. Junker-Frohn, M. Delaney, M.D. Olave, A. Burkart, H. Jaenicke, O. Müller, U. Rascher, and R. Roscher. Growliflower: An image time series dataset for growth analysis of cauliflower. *Journal of Field Robotics (JFR)*, 40(2):173–192, 2022.

[14] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic Segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] C. Lehnert, C. McCool, Y. Sa, and T. Perez. Performance improvements of a sweet pepper harvesting robot in protected cropping environments. *Journal of Field Robotics (JFR)*, 37(7):1197–1223, 2020.

[16] B. Li, J. Lecourt, and G. Bishop. Advances in non-destructive early assessment of fruit ripeness towards defining optimal time of harvest and yield prediction—a review. *Plants*, 7(1):3, 2018.

[17] K. Lin, L. Wang, K. Luo, Y. Chen, Z. Liu, and M.T. Sun. Cross-domain complementary learning using pose for multi-person part segmentation. *IEEE Trans. on Circuits and Systems for Video Technology (TCSVT)*, 31(3):1066–1078, 2020.

[18] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint*, arXiv:1711.05101, 2017.

[19] P. Lottes, J. Behley, N. Chebrolu, A. Milioto, and C. Stachniss. Joint Stem Detection and Crop-Weed Classification for Plant-specific Treatment in Precision Farming. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2018.

[20] P. Lottes, J. Behley, N. Chebrolu, A. Milioto, and C. Stachniss. Robust joint crop-weed classification and stem detection using image sequences. *Journal of Field Robotics (JFR)*, 37(1):20–34, 2020.

[21] Y. Lu and S. Young. A survey of public datasets for computer vision tasks in precision agriculture. *Computers and Electronics in Agriculture*, 178:105760, 2020.

[22] F. Magistri, Y. Pan, J. Bartels, J. Behley, C. Stachniss, and C. Lehnert. Improving Robotic Fruit Harvesting Within Cluttered Environments Through 3D Shape Completion. *IEEE Robotics and Automation Letters (RA-L)*, 9(8):7357–7364, 2024.

[23] M. Malladi, N. Chebrolu, I. Scacchetti, L. Lobefaro, T. Guadagnino, B. Casseau, H. Oh, L. Freissmuth, M. Karppinen, J. Schweier, S. Leutenegger, J. Behley, C. Stachniss, and M. Fallon. DigiForests: A Longitudinal LiDAR Dataset for Forestry Robotics. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2025.

[24] R. Marcuzzi, L. Nunes, L. Wiesmann, J. Behley, and C. Stachniss. Mask-Based Panoptic LiDAR Segmentation for Autonomous Driving. *IEEE Robotics and Automation Letters (RA-L)*, 8(2):1141–1148, 2023.

[25] E. Marks, J. Bömer, F. Magistri, A. Sah, J. Behley, and C. Stachniss. BonnBeetClouds3D: A Dataset Towards Point Cloud-Based Organ-Level Phenotyping of Sugar Beet Plants Under Real Field Conditions. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2024.

[26] E. Marks, M. Sodano, F. Magistri, L. Wiesmann, D. Desai, R. Marcuzzi, J. Behley, and C. Stachniss. High precision leaf instance segmentation for phenotyping in point clouds obtained under real field conditions. *IEEE Robotics and Automation Letters (RA-L)*, 8(8):4791–4798, 2023.

[27] L. McInnes, J. Healy, and S. Astels. HDBSCAN: Hierarchical sensity based clustering. *Journal of Open Source Software*, 2(11):205–206, 2017.

[28] A. Milioto, J. Behley, C. McCool, and C. Stachniss. LiDAR Panoptic Segmentation for Autonomous Driving. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020.

[29] A. Milioto, P. Lottes, and C. Stachniss. Real-time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2018.

[30] H.A. Montes, J. Le Louedec, G. Cielniak, and T. Duckett. Real-time detection of broccoli crops in 3d point clouds for autonomous robotic harvesting. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020.

[31] D. Neven, B.D. Brabandere, M. Proesmans, and L. Van Gool. Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Bandwidth. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[32] Y. Pan, F. Magistri, T. Läbe, E. Marks, C. Smitt, C. McCool, J. Behley, and C. Stachniss. Panoptic Mapping with Fruit Completion and Pose Estimation for Horticultural Robots. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.

[33] I. Pérez-Borrero, D. Marín-Santos, M.E. Gegúndez-Arias, and E. Cortés-Ancos. A fast and accurate deep learning method for strawberry instance segmentation. *Computers and Electronics in Agriculture*, 178:105736, 2020.

[34] G. Roggiolani, M. Sodano, T. Guadagnino, F. Magistri, J. Behley, and C. Stachniss. Hierarchical Approach for Joint Semantic, Plant Instance, and Leaf Instance Segmentation in the Agricultural Domain. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2023.

[35] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proc. of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[36] D. Schunck, F. Magistri, R.A. Rosu, A. Cornelißen, N. Chebrolu, S. Paulus, J. Léon, S. Behnke, C. Stachniss, H. Kuhlmann, and L. Klingbeil. Pheno4d: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. *Plos One*, 16(8):e0256340, 2021.

[37] K. Sirohi, R. Mohan, D. Büscher, W. Burgard, and A. Valada. EfficientLPS: Efficient LiDAR Panoptic Segmentation. *IEEE Trans. on Robotics (TRO)*, 38(3):1894–1914, 2021.

[38] C. Smitt, M. Halstead, P. Zimmer, T. Läbe, E. Guclu, C. Stachniss, and C. McCool. PAg-NeRF: Towards fast and efficient end-to-end panoptic 3D representations for agricultural robotics. *IEEE Robotics and Automation Letters (RA-L)*, 9(1):907–914, 2024.

[39] C. Smitt, M. Halstead, T. Zaenker, M. Bennewitz, and C. McCool. PATHoBot: A robot for glasshouse crop phenotyping and intervention. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2021.

[40] B. Triggs, P.F. McLauchlan, R.I. Hartley, and A.W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Proc. of the Intl. Workshop on Vision Algorithms: Theory and Practice*, 1999.

[41] T. van Klompenburg, A. Kassahun, and C. Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177:105709, 2020.

[42] J. Weyler, F. Magistri, E. Marks, Y.L. Chong, M. Sodano, G. Roggiolani, N. Chebrolu, C. Stachniss, and J. Behley. Phenobench: A large dataset and benchmarks for semantic image interpretation in the agricultural domain. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(12):9583–9594, 2024.

[43] J. Weyler, J. Quakernack, P. Lottes, J. Behley, and C. Stachniss. Joint plant and leaf instance segmentation on field-scale uav imagery. *IEEE Robotics and Automation Letters (RA-L)*, 7(2):3787–3794, 2022.

[44] X. Xu, P. Gao, X. Zhu, W. Guo, J. Ding, C. Li, M. Zhu, and X. Wu. Design of an integrated climatic assessment indicator (icai) for wheat production: A case study in jiangsu province, china. *Ecological Indicators*, 101:943–953, 2019.

[45] Z. Zhou, Y. Zhang, and H. Foroosh. Panoptic-PolarNet: Proposal-Free LiDAR Point Cloud Panoptic Segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[46] J. Zhu, R. Zhai, H. Ren, K. Xie, A. Du, X. He, C. Cui, Y. Wang, J. Ye, J. Wang, X. Jiang, Y. Wang, C. Huang, and W. Yang. Crops3D: a diverse 3D crop dataset for realistic perception and segmentation toward agricultural applications. *Scientific Data*, 11, 2024.