This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

AppleGrowthVision: A large-scale stereo dataset for phenological analysis, fruit detection, and 3D reconstruction in apple orchards

Laura-Sophia von Hirschhausen[†‡]
Fraunhofer HHI
vonhirschhausen@campus.tu-berlin.de

Jannes S. Magnusson[‡]
Fraunhofer HHI
jannes.magnusson@hhi.fraunhofer.de

Mykyta Kovalenko[‡]
Fraunhofer HHI
mykyta.kovalenko@hhi.fraunhofer.de

Fredrik Boye[‡]
Fraunhofer IVI
fredrik.boye@ivi.fraunhofer.de

Tanay Rawat
Fraunhofer IVI
tanay.rawat@ivi.fraunhofer.de

Peter Eisert
Fraunhofer HHI
peter.eisert@hhi.fraunhofer.de

Anna Hilsmann
Fraunhofer HHI
anna.hilsmann@hhi.fraunhofer.de

Sebastian Pretzsch
Fraunhofer IVI
sebastian.pretzsch@ivi.fraunhofer.de

Sebstian Bosse
Fraunhofer HHI
sebastian.bosse@hhi.fraunhofer.de

Abstract

Deep learning has transformed computer vision for precision agriculture, yet apple orchard monitoring remains limited by dataset constraints. The lack of diverse, realistic datasets and the difficulty of annotating dense, heterogeneous scenes. Existing datasets overlook different growth stages and stereo imagery, both essential for realistic 3D modeling of orchards and tasks like fruit localization, yield estimation, and structural analysis. To address these gaps, we present AppleGrowthVision, a large-scale dataset comprising two subsets. The first includes 9,317 high resolution stereo images collected from a farm in Brandenburg (Germany), covering six agriculturally validated growth stages over a full growth cycle. The second subset consists of 1,125 densely annotated images from the same farm in Brandenburg and one in Pillnitz (Germany), containing a total of 31,084 apple labels. AppleGrowthVision provides stereo-image data with agriculturally validated growth stages, enabling precise phenological analysis and 3D reconstructions. Extending MinneApple with our data improves YOLOv8 performance by 7.69 % in terms of F1-score, while adding it to MinneApple and MAD boosts Faster R-CNN F1-score by 31.06 %. Additionally, six BBCH stages were predicted with over 95 % accuracy using VGG16, ResNet152, DenseNet201, and MobileNetv2. AppleGrowthVision bridges the gap between agricultural science and computer vision, by enabling the development of robust models for fruit detection, growth modeling, and 3D analysis in precision agriculture. Future work includes improving annotation, enhancing 3D reconstruction, and extending multimodal analysis across all growth stages.

1. Introduction

Agricultural digitalization demands sophisticated datasets to develop advanced decision support systems and automate labor intensive tasks such as fruit picking. Monitoring apple orchards presents unique computational challenges due to highly variable environmental conditions—fluctuating light, unpredictable weather patterns, and dramatic seasonal transformations—all of which significantly impact image quality and consistency. These constraints have severely

[†]This research has received funding from the German Federal Ministry for Economic Affairs and Climate Action as part of the NaLamKI project under Grant 01MK21003D.

[‡]Equal contribution.

limited the development of robust models capable of reliable crop management and surveillance across diverse conditions [9, 18, 26].

Large-scale, temporally diverse orchard imagery offers transformative potential for precision agriculture. Such datasets significantly enhance yield estimation by enabling computer vision models to learn intricate fruit development patterns, effectively manage complex occlusions between densely clustered apples, and generate statistically reliable predictions, directly addressing persistent challenges in resource allocation and post-harvest waste [11, 15, 37]. Furthermore, comprehensive visual data improves model adaptability to dynamic field conditions, substantially enhancing algorithmic robustness and cross-orchard generalization while minimizing costly recalibration requirements [21].

We introduce AppleGrowthVision, the first publicly available dataset capturing apple orchards throughout a complete phenological cycle with calibrated stereo imagery. The dataset is publicly available at fraunhofer-hhi.github.io/AppleGrowthVision. Our dataset uniquely provides scientifically validated agricultural annotations according to the "Biologische Bundesanstalt, Bundessortenamt, und Chemische Industrie" (BBCH) scale [10], documenting six critical growth stages of apple trees. The entire Brandenburg collection comprises precisely calibrated stereo image pairs, enabling accurate three-dimensional analysis and reconstruction. Our contributions include:

- A comprehensive, multi-site dataset integrating orchards from different regions and complementary existing datasets, demonstrating quantifiable improvements of up to 28 % in apple detection accuracy and count estimation.
- State-of-the-art classification models fine-tuned specifically for precise BBCH growth stage prediction in apple trees, achieving over 95% accuracy across growth stages and enabling data-driven phenological monitoring and intervention planning.
- Novel computational techniques for reconstructing complex, large-scale orchard environments in three dimensions, supporting structural analysis, pruning optimization, and precisely targeted agricultural interventions.

2. Related Work

Fruit detection in orchard environments has advanced significantly with the several specialized apple tree datasets. The MinneApple dataset provides 1,000 high-resolution images with over 41,000 annotated apple instances using polygonal masks to enhance object detection, segmentation, and fruit counting in orchards. This dataset supports patch-based fruit clustering analysis to facilitate yield estimation [11]. The dataset from Washington State University consists of 2,298 images from robotic harvesting studies, featuring multiple apple varieties under different lighting conditions in modern fruiting wall architectures [1]. Michigan State University's apple detection dataset comprises of 1,246 RGB images of apple trees with varied lighting and occlusion conditions, developed alongside the Occluder-Occludee Relational Network (O2RNet), for detecting clustered apples [3]. The recent Monastery Apple Dataset (MAD) introduced a semi-supervised dataset comprising 5,545 images collected via drone from 16 apple trees in a German monastery, including 105 labeled images to improve detection across diverse lighting conditions [15].

BBCH first digit	principal stage
0	Bud development
1	Leaf development
3	Shoot development
5	Inflorescence emergence
6	Flowering
7	Development of fruit
8	Maturity of fruit and seed
9	Senescence, beginning of dormancy

Table 1. Used BBCH Principal growth stages for pome fruit [36]
.

While these datasets have advanced fruit detection capabilities, they lack critical growth stage annotations and typically cover limited time frames within a single growing season. AppleGrowthVision adresses these limitations by capturing an entire growth cycle with expert-validated phenological stages, enabling comprehensive analysis of fruit development and detection under real-world conditions.

Phenology serves as a fundamental metric for plant classification and comparison. The extended BBCH scale represents a standardized system for describing vegetative stages in both research and agricultural management. Initially introduced by Hacke et al. in 1992 [10] and later extended for apple trees by Meier in 1994 [36], this scale has been adopted for numerous plant species [22]. The BBCH system employs a two-digit coding system ranging from 0 to 99, where the first digit indicates the principal growth stage (Table 1), and the second digit defines secondary stages. This system's key advantage is its consistent representation of similar phenological stages across diverse plant species.

Recent work by Nguyen et al. [25] focused on annotating apple tree data for two principal growth stages and fine-tuneing multiple network architectures for classification.

Three-dimensional reconstruction of orchard environments presents substantial technical challenges. Current approaches typically operate under controlled laboratory conditions to eliminate wind effects and scene complexity [4, 8], or focus exclusively on leafless trees to simplify the reconstruction task[39]. CherryPicker [23] require up

to 250 images per tree making detailed reconstruction of extensive orchards computationally prohibitive and operationally impractical.

In contrast, AppleGrowthVision demonstrates that using a calibrated stereo camera setup dramatically improves reconstruction efficiency, enabling detailed 3D modeling of large orchard scenes with just 18 images per tree. This approach makes comprehensive digital twins of commercial orchards feasible for the first time, supporting advanced structural analysis and precision agricultural interventions at scale.

3. The AppleGrowthVision Dataset

In response to these limitations, we introduce the AppleGrowthVision Dataset. A comprehensive, longitudinal resource that captures one growth cycle of apple trees. Unlike previous datasets, AppleGrowthVision includes high-resolution images for robust fruit detection and integrates expert-validated phenological data. This integration provides a dynamic view of apple development across different stages, bridging the gap between static image collections and the temporal complexity of real-world orchard environments. By allowing analyses that account for seasonal changes and vegetative development, our dataset offers a more holistic basis to advance both detection algorithms and agronomic insights.

3.1. Data collection

At orchards of Berlin-Brandenburg Obst (BB Obst), we collected images of the same 33 Jonagold trees on 18 occasions throughout 2022. BB Obst cultivates fruits, is a fruit retailer and is located in Wesendahl near Berlin. Next to cherries, pears, plums, and strawberries many apple varieties are grown in spindle form. We constructed a stereo imaging setup using two Canon EOS 550D cameras with 20mm lenses, which were triggered simultaneously (see figure 1). This setup is calibrated using the calibration object and method described in [6]. Each tree was captured from both sides of the row including front, left, right, and steep angled views, as well as a circular movement around the first tree of the row (Figure 2). Some images were taken in windy conditions, which introduced tree deformations.

Another part of the dataset was collected at the ”Sächsisches Landesamt für Umwelt, Landwirtschaft und Geologie” (LfULG) orchard, located in Pillnitz near Dresden. The orchard is used for fruit cultivation research and educational purposes. Many different apple varieties are grown in rows of about 70m length, with trees spaced 1m apart in spindle form and 3m between rows. For a subset of trees, six images per tree were taken from different angles, as described in the previous section using a smartphone camera. Each tree is referenced by row number and tree position within the row. Data collection took place



Figure 1. Stereo camera setup with two calibrated Canon EOS 550D.

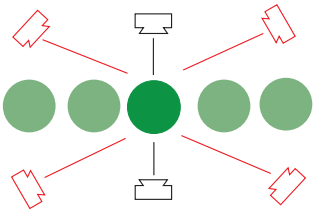


Figure 2. Positions per tree in a row on an orchard using a stereo setup per position

at various dates in 2023, covering stages from blossom to ripe apples. In 2024, the same trees were documented once more, specifically with the apple count. Additional data was collected from other rows, capturing two images per tree using a smartphone. These images were taken by walking along the rows in one direction and back on the other side, ensuring that the first and last images in the subset correspond to the front and back views of the first tree. In total, the Dresden subset consist of 669 labeled images. In 2024, the orchard experienced a significant loss in fruit due to late frost.

For 10 trees which are included in the images taken in Pillnitz in 2024, a manual reference count for the number of apples on the trees was performed. It was done without harvest, all apples remained on the trees. Information about the counted number of apples is provided with the dataset.

3.2. Data Annotation

The dataset was annotated for apple recognition using an in-house developed annotation software, designed to streamline the labeling process through a combination of AI-assisted automation and human verification. To facilitate annotation, we trained YOLOv8 on a manually labeled subset of 108 images in non-stereo and 70 stereo images, combined with the full MinneApple dataset. This initial model generated automatic annotations for the remaining, unlabeled images. The annotations consist of bounding boxes drawn around individual apples and were created only for images representing BBCH growth stages five to nine, as these stages are most relevant for fruit visibility and detection.



Figure 3. Data acquisition in Pillnitz

To improve the quality of the annotations, human annotators reviewed and corrected the AI-generated labels for 70 images of the stereo-format, and for 777 of the non-stereo images of the dataset. The final annotations consist of bounding boxes around apples in the images, stored in Darknet/YOLO format used by the Ultralytics model training framework. This semi-automated annotation approach significantly reduced the time and effort required for labeling while maintaining a high level of accuracy.

3.3. Growth Stage Classification

From each date in the dataset, a subset of images was randomly selected for growth stage categorization based on the extended BBCH scale for pome fruit. This classification was performed manually by an expert from LfULG Sachsen. Local variations within different sections of the orchard and variability in individual tree development result in different secondary growth stages of trees outside the random sample. Table 2 shows an overview of all stages contained in the AppleGrowthVision dataset, sorted by scale. This information allows for the selection of subsets of the dataset, if a specific growth stage is of special interest.

Figure 4 shows images from each of the six stages of the Brandenburg subset of the AppleGrowthVision dataset.

4. Dataset Evaluation

We conducted a series of experiments aimed at assessing the dataset's effectiveness in two key areas. Training and validating object detection models for fruit localization and growth stage prediction, and enabling 3D reconstruction of orchard scenes from calibrated stereo imagery. These evaluations provide insight into the dataset's quality, robustness, and relevance for advancing both phenological monitoring and spatial analysis in orchard environments.

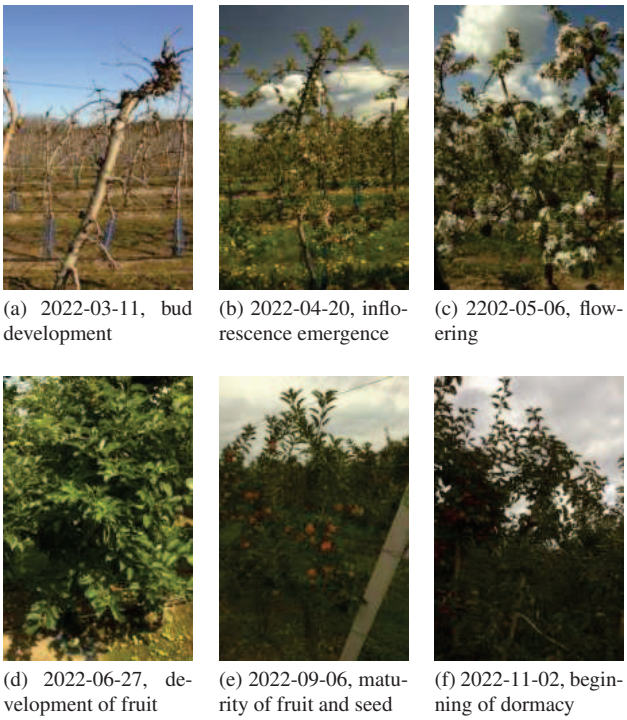
BBCH	date	location
01	22.03.2023	BB Obst
03	11.03.2022	BB Obst
	25.03.2022	BB Obst
54	01.04.2022	BB Obst
55	08.04.2022	BB Obst
56	21.04.2023	BB Obst
57	29.04.2022	BB Obst
60	04.05.2023	BB Obst
61/65	25.04.2023	Pillnitz
65	04.05.2023	Pillnitz
	06.05.2022	BB Obst
67	13.05.2022	BB Obst
	16.05.2023	BB Obst
71/72	24.05.2023	Pillnitz
72	30.05.2022	BB Obst
73	13.06.2022	BB Obst
74	27.06.2022	BB Obst
77	03.08.2022	BB Obst
81	23.08.2023	Pillnitz
85	26.08.2024	Pillnitz
87	06.09.2022	BB Obst
	19.09.2023	Pillnitz
91	02.11.2022	BB Obst

Table 2. extended BBCH stages included in dataset

4.1. Apple Detection: YOLOv8, Faster R-CNN

We evaluated the impact of our dataset on improving fruit detection models in orchard environments. We used YOLOv8 and Faster R-CNN [27], two widely adopted object detection architectures, to assess how different training configurations affect performance. For further evaluation, applying a transformer-based approach would be valuable, as transformer architectures have shown promising results when using large amounts of image data [2, 5] and could enhance detection accuracy in dense, occluded orchard scenes.

Our evaluation focused on how extending publicly available datasets like MinneApple and MAD improves fruit detection accuracy. The improvement in apple detection accuracy is assessed on the MAD and MinneApple validation sets. The models were trained on various dataset combinations and evaluated using Average Precision (AP), Average Recall (AR), mean Average Precision at IoU thresholds of 0.5 (mAP@0.5) and 0.5:0.95 (mAP@0.5:0.95), and F1 score.



(a) 2022-03-11, bud development
(b) 2022-04-20, inflorescence emergence
(c) 2202-05-06, flowering
(d) 2022-06-27, development of fruit
(e) 2022-09-06, maturity of fruit and seed
(f) 2022-11-02, beginning of dormacy

Figure 4. Example images for each existing growth stage in BBCH scale notation from the dataset from growth cycle 2022 of the Brandenburg subset.

Table 3 summarizes the results of our experiments. We observed that augmenting the MinneApple dataset with AppleGrowthVision training data, which contains manually optimized annotations, improves detection performance of YOLOv8 by 5.87 % in terms of F1-score. Furthermore, incorporating the MAD dataset enhances model performance by an additional 1.82 %, demonstrating the benefits of dataset diversity and additional training data. Training solely on MinneApple resulted in the lowest performance across four of the five metrics when compared to configurations that included all three datasets, indicating limited generalization to more diverse data. The best results relative to training on MinneApple alone were achieved when all three datasets were combined. This configuration yielded the highest AR- and mAP-values, and F1-score, highlighting the value of combining multiple complementary datasets to build robust object detection models for real-world orchard environments.

The experiments with Faster R-CNN followed a similar structure but revealed important differences in behavior. When trained on MinneApple and MAD, Faster R-CNN achieved an AP-, AR-, mAP- and F1-scores that were lower than those of YOLOv8 under the same conditions. Adding AppleGrowthVision to the training set improved AP-, AR- and F1-score, demonstrating that Faster R-CNN also benefits from dataset diversity. Interestingly, the best performance across three of the five metrics was achieved when trained solely on MinneApple and AppleGrowthVision, while the worst results occurred when training only on MinneApple and MAD. This indicates higher sensitivity of Faster R-CNN to scene variability and potential overfitting when trained on a limited combination of datasets. Notably, performance recovered when all three datasets were combined in experiment (2): AP increased by 31.74 %, AR by 30.38 %, mAP@0.5 by 42.12 %, mAP@.5:.95 by 57.21 %, and F1 by 31.06 % compared to experiment (3). Although this was still lower than YOLOv8 in the same setup, it confirms that dataset richness contributes positively to generalization, even for models that are more sensitive to training-test domain shifts.

Model	Metric	(1)	(2)	(3)	(4)
YOLOv8	AP	.814	**.822**	.81	.811
	AR	.637	.701	.716	**.733**
	mAP@.5	.753	.793	.805	**.81**
	mAP@.5:.95	.424	.453	.456	**.464**
	F1	.715	.757	.76	**.77**
Faster R-CNN	AP	.766	**.789**	.564	.743
	AR	.613	.639	.497	**.648**
	mAP@.5	.694	**.721**	.47	.668
	mAP@.5:.95	.352	.345	.194	.305
	F1	.681	**.706**	.528	**.692**

Table 3. Comparison of model performance using different dataset combinations: (1) MinneApple, (2) MinneApple + AppleGrowthVision, (3) MinneApple + MAD, (4) MinneApple + AppleGrowthVision + MAD.

These findings suggest that adding diverse and phenologically rich datasets significantly improves fruit detection performance. While YOLOv8 consistently outperforms Faster R-CNN in overall accuracy and robustness, both architectures benefit from integrating multiple datasets. This underscores the importance of leveraging complementary data sources to build resilient object detection models for complex orchard environments.

4.2. BBCH-scale classification

Following the experiments from [25], we finetuned a VGG16 [32], a ResNet152 [12], a DenseNet201 [14], and a MobileNetv2 [28], which were all pretrained on ImageNet1k [38]. We measured the performance in accuracy, precision, recall, and F1 score. Our data was split randomly using 80% training, 10% validation, and 10% test data. Table 4 shows that each method is able to easily distinguish between each principal BBCH stage. A direct comparison to [25] is not applicable since they focused on secondary BBCH growth stages. Nevertheless, several agricul-

Model	Accuracy	Precision	Recall	F1 Score
VGG16	0.988	0.988	0.988	0.988
ResNet152	0.997	0.997	0.997	0.997
DenseNet201	0.999	0.999	0.999	0.999
MobileNetv2	0.999	0.999	0.999	0.999

Table 4. Comparison of model performance on test data for BBCH stage classification.

tural tasks rely on the fine-grained resolution provided by the secondary BBCH growth stages. For example, fungicide application schedules are strongly tied to phenological stages such as petal fall, which is the last stage of flower development [7]. Similarly, accurate yield estimation depends on distinguishing how far fruit development has progressed within BBCH stage 7, as early fruit set (e.g., BBCH 71) and near-mature fruit (e.g., BBCH 77) indicate very different outcomes. Finally, determining optimal harvest time often depends on secondary stages like BBCH 85, which marks full ripeness, rather than simply recognizing that the fruit is in the ripening phase [36].

4.3. Multi-View Stereo Reconstructions

Calibrated stereo vision data is used to extract 3D reconstructions of scenes. Reconstructions enable automated phenological measures and growth monitoring. These applications depend on the number of captured images per scene and the absence of motion for the target objects. Wind that bends trees is one big challenge in outside scenes. For CherryPicker 250 images were used in almost windless conditions to retrieve point clouds for one specific tree [23]. For common representations, such as point clouds [23], NeRFs [24], and Gaussian Splattings [16], it is assumed that camera positions are given. With our calibrated data, state-of-the-art methods like COLMAP [31] and Mast3r [17] are not able to locate the cameras in these large and complex scenes. SIFT [20], used by COLMAP, is not capable to describe features meaningfully to distinct the feature points of orchard scenes. Instead we extracted DISK [35] features, matched them with LightGlue [19] within the HLOC framework [29, 30], and applied low level adjustments in COLMAP to reconstruct some of these large scenes. In COLMAP we provided an initial pair using the extrinsic calibration priors of one image pair. Then we reconstructed the scene on the initial pair and only the left camera images, adding the right camera images in a second step to optimize computation speed. These reconstructions show that reconstructions of orchards with only stereo images is possible and serve as a baseline for further research on reconstructing these challenging scenes to derive point clouds, NeRFs, and Gaussian Splattings. One reconstructed scene is shown in figure 5.



Figure 5. Example of a reconstructed orchard scene of apple trees from BB Obst.

5. Conclusion and Discussion

This work presents AppleGrowthVision, a novel dataset designed to advance computer vision and 3D applications in precision agriculture, particularly for fruit detection and phenological analysis in apple orchards. Existing datasets, such as MinneApple and MAD, have contributed significantly to orchard-based computer vision research but lack agriculturally validated growth stages and comprehensive 3D reconstruction capabilities. Our dataset fills this gap by providing high-resolution stereo images collected over an entire growth cycle with expert-validated BBCH growth stages, enabling more precise modeling of fruit development.

To begin a simple initial evaluation of AppleGrowthVision's effectiveness, we conducted eight experiments using YOLOv8 and Faster R-CNN, comparing different training configurations to assess improvements in fruit detection performance. The results show that integrating AppleGrowthVision into training enhances detection accuracy across multiple performance metrics. Secondly, we fine-tuned state-of-the-art classification models to predict BBCH growth stages, enabling high-precision support for decision-making and planning in agriculture. All growth stages present in our dataset can be predicted with an accuracy exceeding 95 %. Thirdly, we used calibrated stereo vision data to reconstruct 3D orchard scenes, supporting automated phenological monitoring despite challenges such as wind and dense vegetation. Since traditional feature matching methods like SIFT failed in these complex environments, we implemented a custom pipeline using DISK features, LightGlue, and COLMAP with calibration priors. This approach successfully produced multi-view stereo reconstructions, providing a baseline for future work on point clouds, NeRFs, and Gaussian Splattings.

While AppleGrowthVision represents an advancement in orchard-based computer vision research, several challenges remain. Automated annotation is not yet sufficient for reliable crop monitoring and yield prediction for the entire data set. Therefore, further improvements in automatic labeling techniques are necessary to enhance annotation accuracy and consistency across the dataset. On the one hand,

many agricultural tasks depend on the detailed resolution of secondary BBCH growth stages. For instance, fungicide application is timed to stages like petal fall [7], yield estimation varies between early (BBCH 71) and late fruit set (BBCH 77), and optimal harvest decisions rely on stages like BBCH 85, which marks full ripeness [36]. Therefore, a fully annotated dataset of secondary growth stages is essential for developing reliable and useful models in precision agriculture. On the other hand, further research is needed to determine the level of annotation accuracy—e.g., benchmark F1-scores—required for primary growth stages to develop reliable agricultural technology. This would also help assess whether the annotations of AppleGrowthVision meet those requirements.

To elevate AppleGrowthVision to a widely used benchmark dataset, further validation is needed using both high-capacity and efficient models. Transformer-based architectures such as DETR [2] or Vision Transformers (ViT) [5] have demonstrated strong performance in modeling complex spatial relationships, particularly when trained on large-scale datasets. Applying such models to AppleGrowthVision could improve detection accuracy in dense, occluded orchard scenes and provide insights into scalability across large and variable datasets. At the same time, ensuring real-world usability requires evaluating lightweight models that enable real-time inference on edge devices. YOLOv8 shows promising results in this regard, balancing accuracy and efficiency. However, further benchmarking with resource-aware models such as YOLO-NAS, MobileNetV3, or EfficientDet [13, 33, 34] is necessary to assess performance in constrained environments typical for agricultural robotics and embedded systems.

Despite the benefits of stereo imagery, handling occlusions, scale variations, and fine-grained structure details remains challenging. Additionally, data collection or synthetic augmentation is needed to ensure complete coverage of all BBCH growth stages, including substages, for a more holistic phenological analysis. Finally, combining RGB, multispectral, and depth information could further improve object detection, classification, and phenotypic trait analysis in orchard environments. With further research and refinement, AppleGrowthVision has the potential to serve as a benchmark dataset for fruit detection, growth modeling, and 3D reconstruction in orchard environments, laying the foundation for future advancements in precision agriculture and autonomous farming technologies.

References

[1] Santosh Bhusal, Manoj Karkee, and Qin Zhang. Apple dataset benchmark from orchard environment in modern fruiting wall, 2019. Open-access dataset. 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 4, 7

[3] Pengyu Chu, Zhaojian Li, Kaixiang Zhang, Dong Chen, Kyle Lammers, and Renfu Lu. O2rnet: Occluder-occludee relational network for robust apple detection in clustered orchard environments. *arXiv preprint arXiv:2303.04884*, 2023. 2

[4] Martin Churuvija, Ranjan Sapkota, Dawood Ahmed, and Manoj Karkee. A pose-versatile imaging system for comprehensive 3D modeling of planar-canopy fruit trees for automated orchard operations. *Computers and Electronics in Agriculture*, 230:109899, 2025. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 4, 7

[6] Peter Eisert. Model-Based Camera Calibration Using Analysis by Synthesis Techniques. In *VMV*, pages 307–314, 2002. 3

[7] M. A. Ellis, D. C. Ferree, R. C. Funt, and L. V. Madden. Effects of an apple scab-resistant cultivar on use patterns of inorganic and organic fungicides and economics of disease control. *Plant Disease*, 82(4):428–433, 1998. 6, 7

[8] Tian Gao, Feiyu Zhu, Puneet Paul, Jaspreet Sandhu, Henry Akrofi Doku, Jianxin Sun, Yu Pan, Paul Staswick, Harkamal Walia, and Hongfeng Yu. Novel 3D Imaging Systems for High-Throughput Phenotyping of Plants. *Remote Sensing*, 13(11):2113, 2021. 2

[9] Michael Gerstenberger, Mykyta Kovalenko, David Przewozny, Jannes Magnusson, Eike Gassen, Jakub Pawlak, Jochen Hirth, Laura von Hirschhausen, Detlef Runde, Anna Hilsmann, Peter Eisert, and Sebastian Bosse. Automatisierte frucht-und pflanzenerkennung in apfelplantagen durch künstliche intelligenz. In *44. GIL-Jahrestagung, Biodiversität fördern durch digitale Landwirtschaft*, pages 59–70. Gesellschaft für Informatik eV, 2024. 2

[10] Hack. H., Bleiholder L., Meier U., Schnockfricke U., Weber E., and Witzenberger A. Einheitliche codierung der phänologischen entwicklungsstadien von kultur- und schadpflanzen- erweiterte bbch-skala. *Nachrichtenblatt des Deutschen Pflanzenschutzdienstes*, 44:265–270, 1992. 2

[11] Nicolai Häni, Pravakar Roy, and Volkan Isler. Minneapple: A benchmark dataset for apple detection and segmentation. *arXiv preprint arXiv:1909.06441*, 2020. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[13] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3, 2019. 7

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 5

[15] Robert Johanson, Christian Wilms, Ole Johannsen, and Simone Frintrop. S3ad: Semi-supervised small apple detection in orchard environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, United States, 2024. IEEE. Equal contribution by Robert Johanson and Christian Wilms. 2

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4):1–14, 2023. 6

[17] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 6

[18] Yang Li and Xuewei Chao. Toward sustainability: Trade-off between data quality and quantity in crop pest recognition. *Frontiers in Plant Science*, Volume 12 - 2021, 2021. 2

[19] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17581–17592, 2023. ISSN: 2380-7504. 6

[20] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1150–1157 vol.2, Kerkyra, Greece, 1999. IEEE. 6

[21] Yuzhen Lu and Sierra Young. A survey of public datasets for computer vision tasks in precision agriculture. *Computers and Electronics in Agriculture*, 178:105760, 2020. 2

[22] Uwe Meier, Hermann Bleiholder, Liselotte Buhr, Carmen Feller, Helmut Hack, Martin Heß, Peter D. Lancashire, Uta Schnock, Reinhold Stauß, Theo van den Boom, Elfriede Weber, and Peter Zwerger. The bbch system to coding the phenological growth stages of plants – history and publications. *Journal für Kulturpflanzen*, 61(2), 2009. 2

[23] Lukas Meyer, Andreas Gilson, Oliver Scholz, and Marc Stamminger. CherryPicker: Semantic Skeletonization and Topological Reconstruction of Cherry Trees. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6244–6253, Vancouver, Canada, 2023. 2, 6

[24] Lukas Meyer, Andreas Gilson, Ute Schmidt, and Marc Stamminger. Fruitnerf: A unified neural radiance field based fruit counting framework. In *IROS*, 2024. 6

[25] Xuan Khanh Nguyen, Bastian Braun, Nico Heider, and Martin Schieck. Recognition of phenological development stages of apple blossoms using computer vision. In *45. GIL-Jahrestagung, Digitale Infrastrukturen für eine nachhaltige Land-, Forst- und Ernährungswirtschaft*, pages 119–130. Gesellschaft für Informatik e.V., Bonn, 2025. 2, 5

[26] Lorenz Gunreben Nico Heider. A survey of datasets for computer vision in agriculture. *Proceedings of the 45th GIL Annual Conference (GIL-Jahrestagung): Digitale Infrastrukturen für eine nachhaltige Land-, Forst- und Ernährungswirtschaft*, 2025. 2

[27] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015. 4

[28] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 5

[29] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 6

[30] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 6

[31] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[32] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, 2015. arXiv: 1409.1556. 5

[33] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection, 2020. 7

[34] Juan R. Terven and Diana M. Cordova-Esparza. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine Learning and Knowledge Extraction*, 5(4):83, 2024. 7

[35] MichałTyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. In *Advances in Neural Information Processing Systems*, pages 14254–14265. Curran Associates, Inc., 2020. 6

[36] Meier U., Graf H., Hack H., Hess M., Kennel W., Klose R., Mappes D., Seipp D., Strauss R. Streif, and J. Van den Boom. Phänologische entwicklungsstadien des kernobstes (malus domestica borkh. und pyrus communis l.), des steinobstes (prunus-arten), der johannisbeere (ribes-arten) und der erdbeere (fragaria ananassa duch.). *Nachrichtenblatt des Deutschen Pflanzenschutzdienstes*, 46:141–153, 1994. 2, 6, 7

[37] Christian Wilms, Robert Johanson, and Simone Frintrop. Localizing small apples in complex apple orchard environments. *arXiv preprint arXiv:2202.11372*, 2022. 2

[38] Yang You, Zhao Zhang, Cho-Jui Hsieh, James Demmel, and Kurt Keutzer. ImageNet Training in Minutes. In *Proceedings of the 47th International Conference on Parallel Processing*, New York, NY, USA, 2018. Association for Computing Machinery. event-place: Eugene, OR, USA. 5

[39] Xilei Zeng, Hao Wan, Zeming Fan, Xiaojun Yu, and Hengrong Guo. MT-MVSNet: A lightweight and highly accurate convolutional neural network based on mobile transformer for 3D reconstruction of orchard fruit tree branches. *Expert Systems with Applications*, 268:126220, 2025. 2