Toward Multimodal Agricultural Digital Twins: Integrating Autonomous UGV Monitoring, High-Precision Phenotyping, and Diffusion-Based Generative Data Augmentation

Yuichiro Nomura[1], **Hiroshi Mineno**[1,2]

[1] Dept. of Computer Science in the Faculty of Informatics, Shizuoka University, Japan
[2]Research Institute of Green Science and Technology, Shizuoka University, Japan
nomura.yuichiro@inf.shizuoka.ac.jp, mineno.hiroshi@shizuoka.ac.jp

Abstract

Agricultural digital twins require continuous multimodal sensing, robust field robotics, and scalable machine learning models that can be generalized across heterogeneous environments. This paper summarizes a series of practical studies conducted in our laboratory that contribute to building the foundation of an Agricultural Digital Twin (Agri-DT) platform: (i) a low-cost self-charging UGV for autonomous greenhouse monitoring, (ii) high-precision mobile phenotyping techniques for grapevine reproductive organs (inflorescences and flowers), and (iii) a diffusion-model-based generative data augmentation framework enabling domain-adaptive phenotypic analysis with minimal annotation costs. Experimental studies conducted in greenhouses and open-field vineyards showed the potential of our approach to improve autonomous mobile monitoring and plant structure detection tasks through domain-adaptive data generation. Based on these results, we discuss a scalable agricultural digital twin construction method that integrates multimodal data streams (environmental data, growth videos, stress indicators, labor records) to enable crop growth simulation, anomaly prediction, and production planning.

Introduction

Digital Twins serve as virtual representations of physical farming systems, supporting simulation, monitoring, and prediction, and are increasingly viewed as essential components of next-generation smart agriculture. A practical Agricultural Digital Twin (Agri-DT) requires spatial–temporal data capturing plant growth and canopy structure, high-precision phenotyping for structural and reproductive traits, and generalizable perception models capable of training across heterogeneous environments.

However, agricultural environments exhibit extreme variability in factors such as lighting, weather conditions, cultivar differences, and seasonal transitions, which makes robust sensing and modeling challenging. To address these issues, our laboratory has developed the following three complementary technologies. This paper integrates these efforts and outlines a pathway toward multimodal Agri-DTs.

1. Autonomous UGV for continuous greenhouse monitoring (Zobaer et al. 2025).
2. Mobile phenotyping for plant reproductive organs (inflorescences and flowers) (Rahim, Utsumi, and Mineno 2022).
3. Diffusion-based domain-adaptive generative data augmentation for robust phenotypic detection (Hirahara et al. 2025).

Related Work

Autonomous mobile robots are becoming increasingly important in agricultural monitoring because they can repeatedly observe crop canopies in both time and space. Many existing unmanned ground vehicles (UGVs) rely on LiDAR-based SLAM or GNSS for stable navigation (Fasiolo et al. 2023). While these systems achieve excellent accuracy, their high cost and power consumption hinder adoption in greenhouses and small-to-medium-sized farms. Recent research has explored lower-cost alternatives, such as position estimation using reference markers and lightweight depth sensors. However, challenges remain in achieving long-term autonomous operation, including self-charging and continuous patrol cycles. Consequently, there is a growing demand for sustainable, low-cost UGV monitoring systems that function as core data collection devices for agricultural digital twins. This research contributes to this direction by developing an AprilTag-based UGV platform with autonomous charging capabilities. This enables continuous monitoring without human intervention.

Meanwhile, deep learning has advanced agricultural phenotyping, enabling robust detection and segmentation of leaves, fruits, and canopy structures. However, achieving high accuracy in actual field environments remains challenging, particularly for small, densely packed reproductive organs like grape flowers. While fine-grained detection is crucial in the early stages of yield prediction, many existing methods rely on limited training data or controlled imaging environments. Recent vineyard-specific approaches have improved flower and inflorescence detection using patch-based processing, instance segmentation, and artificial lighting, but performance tends to degrade under occlusion,

nighttime conditions, and varietal differences (Palacios et al. 2020). To address these challenges, our group developed a two-stage phenotypic analysis pipeline capable of accurately detecting inflorescences and individual flowers in commercial nighttime vineyards. This enables reliable quantification of reproductive traits and provides essential structural information for data-driven yield modeling within digital twins.

Furthermore, domain variability stemming from differences in lighting, season, cultivated varieties, and imaging devices represents another major challenge in agricultural machine learning. GAN-based augmentation techniques have been used for variability mitigation, but issues remain, such as lack of realism and difficulty maintaining geometric consistency with existing annotations. Diffusion models offer higher fidelity and controllability, but their application in agriculture remains limited (Muhammad et al. 2023), and annotation-preserving generation techniques have been scarcely researched. To bridge this gap, we introduce "D4," a diffusion-based domain-adaptive data augmentation framework that generates high-quality synthetic images while preserving bounding box and keypoint annotations. Through a two-stage fine-tuning process and annotation-conditional generation, D4 enhances model robustness for downstream phenotypic analysis tasks, strengthening the data foundation required for agricultural digital twins.

Early research on Agri-DTs primarily focused on environmental simulation, climate modeling, and resource optimization, often relying on static datasets and sensor networks. While these studies highlight DT's potential for decision support, they generally lack integration with robotic monitoring, high-resolution phenotyping, and generative data modeling, resulting in limited scalability across crops and environments. Only a handful of recent studies have explored multimodal sensing pipelines, and even fewer have demonstrated the continuous, automated data acquisition and reproductive trait quantification essential for yield modeling. This research addresses these challenges by integrating:

1. Monitoring via Autonomous UGVs: Continuous acquisition of spatiotemporal data
2. High-precision reproductive trait phenotyping: Conducted under field conditions
3. Diffusion-based generative extension: Domain-adaptive model training

By integrating robotics, phenotyping, and generative AI, this study provides a unified foundation for a multimodal Agri-DT enabling scalable growth simulation, anomaly detection, and yield prediction. Recent Agri-DT surveys mainly summarize conceptual frameworks and enabling technologies for digital twins in agriculture (Purcell and Neubauer, 2023; Subeesh and Chauhan, 2026). In contrast,

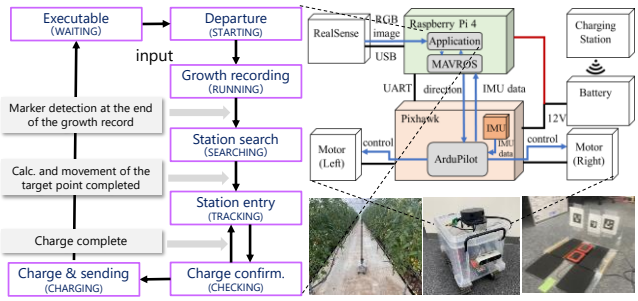

Figure 1: Low-cost autonomous UGV.

our work contributes a deployable multimodal sensing pipeline integrating autonomous UGV monitoring, phenotyping, and generative augmentation that provides the data foundation needed for practical DT applications.

Autonomous UGV for Continuous Greenhouse Monitoring

Building on recent advances in low-cost robotic surveillance technology, our laboratory has developed a low-cost fully autonomous UGV (Zobaer et al. 2025) as shown in Figure 1. This system incorporates Marker-based positioning capabilities and a self-charging mechanism. It achieved a 100% docking success rate across 15 trials with retry function, enabling long-term, continuous spatiotemporal sensing within greenhouse environments. This sustained monitoring capability forms the core of an Agri-DT platform, supporting growth modeling, anomaly detection, and plant observation over several weeks.

The construction of an Agri-DT requires continuous spatiotemporal observation, particularly in greenhouse environments where plant growth progresses dynamically over days to weeks. To support low-cost, sustainable monitoring, we developed an autonomous unmanned ground vehicle (UGV) equipped with self-charging capabilities and low-power sensing modules. This enables uninterrupted, long-term data acquisition without human intervention. The design and functionality of this UGV are based on the system presented in our previous research.

The proposed UGV is built around a Raspberry Pi 4, coordinating sensing, navigation, and power management tasks via ROS. For environmental awareness, the vehicle integrates an Intel RealSense D435 depth camera, used for both obstacle detection and recognition of AprilTag or ArUco fiducial markers. This enables robust visual localization within greenhouse lanes. Low-level driving is handled by a Pixhawk (Cube Orange) controller, achieving precise control of the crawler-based drive system and stability through IMU feedback. A lightweight Pi Camera v2 is mounted for periodic video collection of plant growth.

The system's key features are autonomous docking and self-charging mechanisms, designed for long-term sustained

operation. Battery status is continuously monitored via an LC709203F Li-Po fuel gauge. When the remaining charge falls below the threshold (20%), the UGV is designed to automatically stop video recording and autonomously navigate to a designated charging station. Alignment is performed using depth sensing and Marker-based visual cues, enabling Qi wireless charging. Experimental evaluation confirmed a 100% docking success rate across all 15 trials, including straight, left-angle, and right-angle approaches, utilizing the retry function. Docking was completed within 18 to 35 seconds depending on the approach angle. This demonstrated its exceptional suitability for autonomous operation under energy-constrained conditions.

These capabilities enable the UGV to function as a persistent monitoring platform. It continuously traverses between rows within greenhouses, acquiring time-series RGB-D images while autonomously recharging as needed. The system's low-cost components, modular architecture, and high reliability make it ideal for deployment across multiple greenhouses and integration into large-scale sensing networks. Consequently, the UGV serves as an essential element within the data collection layer of an Agri-DT framework, providing the spatio-temporal information required for growth modeling, anomaly detection, and long-term crop behavior analysis.

Mobile Phenotyping for Plant Reproductive Organs

Accurate quantification of reproductive organs, such as inflorescences and flowers, is essential for early yield estimation and the construction of biologically meaningful Agri-DTs. However, detecting inflorescences and individual flowers in vineyard images in unstructured field environments is challenging due to (i) the extremely small and dense nature of floral structures, (ii) severe occlusion by leaves, and (iii) visual similarity between plant organs. To address these challenges, based on prior research in commercial vineyards, we developed a high-precision phenotyping framework based on a two-stage Mask R-CNN architecture featuring patch-based processing, test-time expansion, and contrast enhancement preprocessing (Rahim, Utsumi, and Mineno 2022) as shown in Figure 2. This illustrates the nighttime vineyard imaging setup used to obtain high-resolution RGB data under controlled artificial illumination.

In the first stage, we detect and segment inflorescences from high-resolution RGB images (6000×4000 pixels) captured at night using a mobile sensing platform equipped with artificial lighting. Patch extraction (2000×1333 pixels) and extensive augmentation processes generate a robust training dataset that transcends varietal and imaging conditions. To ensure high recall, test-time augmentation (TTA) including horizontal flipping and controlled color transformations is

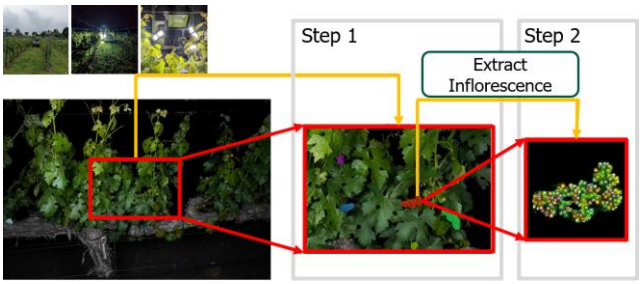

Figure 2: Two-Stage Deep Instance Segmentation.

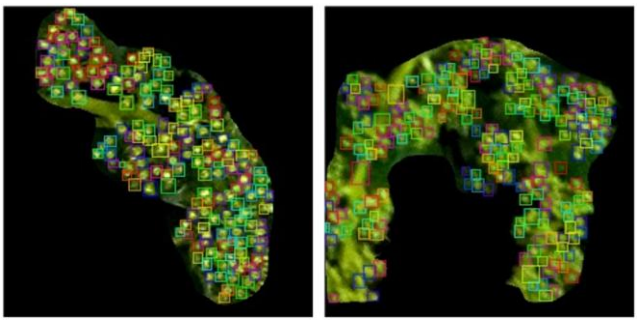

Figure 3: Example of Individual flower detection.

applied during inference. Majority-based instance recognition aggregation recovers missed inflorescences while eliminating false positives. This process achieves a high F1 score of 0.943, surpassing conventional semantic segmentation methods and enabling reliable extraction of reproductive regions even under blurring or partial occlusion.

In the second stage, individual flowers within each segmented inflorescence are detected using a separate Mask R-CNN model. Due to the extremely small size of flower shapes, images are enlarged and normalized before processing, using a sliding window method (256×256 patches) to preserve fine spatial details. To address challenges from low-light and motion blur conditions, contrast enhancement using multiplication and linear contrast operations was introduced as a preprocessing step during inference. Optimal parameters were empirically determined, significantly improving flower detection recall while maintaining high accuracy as shown in Figure 3. This achieved a final flower detection performance of F1 = 0.903, demonstrating high agreement (R^2 = 0.981) with manual ground truth data across three grape varieties. This shows the strong correlation between predicted and ground-truth flower counts, confirming the reliability of our reproductive trait estimation.

This complete pipeline enables accurate quantification of both inflorescences and individual flowers, with an error rate (MAPE < 12%) below the common threshold for plant phenotyping. Furthermore, the system reveals important biological patterns, such as the sensitivity of flower-level detection to shading, providing insights for future multimodal sensing strategies (e.g., multi-view imaging and depth-based reconstruction). As a module within an Agri-DT, this

phenotyping framework contributes reliable reproductive trait measurements essential for predictive yield modeling, growth simulation, and data-driven crop management.

Diffusion-Based Domain-Adaptive Generative Data Augmentation

Deep learning-based plant phenotyping requires large-scale, diverse, and properly annotated datasets. However, collecting and annotating agricultural image data is extremely challenging due to complex canopy structures, shading, environmental variability, and the minute size of key plant organs. Domain shifts caused by variations in lighting conditions (day/night), growth stages, and cultivation systems further degrade model performance and limit generalizability to real-world production environments. To address these challenges, we developed "D4," a diffusion-based text-guided generative data augmentation framework (Hirahara et al. 2025). It enables domain adaptation while preserving bounding box and keypoint annotations for object detection tasks.

The D4 method utilizes a ControlNet-based text-guided diffusion model through a two-stage fine-tuning process as shown in Figure 4. In the first stage, it learns broad image features using large amounts of unlabeled images extracted from UGV-monitored vineyard videos, encoding domain characteristics via Canny edge maps and text prompts. Since no annotations are required, it effectively leverages the abundant real agricultural images obtainable through autonomous robotic sensing. Stage 2 fine-tunes the model using a small annotated dataset of just 50 nighttime images labeled with shoots and nodes. Training employs "annotation plot images" that directly embed bounding boxes and keypoint structures into the input. This enables D4 to learn precise geometric relationships between plant structures and annotation coordinates.

After fine-tuning, D4 can generate domain-adapted annotated images solely through prompt conditioning. For example, it can convert nighttime annotated images into daytime

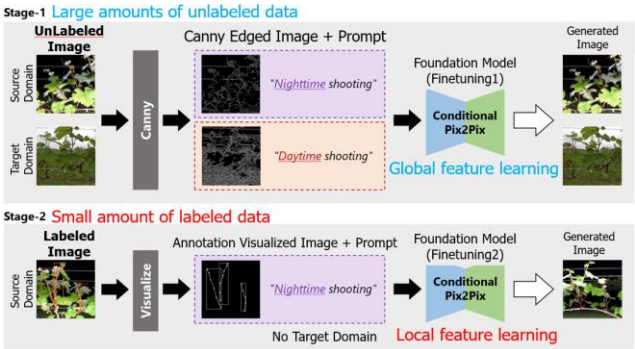

Figure 4: Diffusion-based domain-adaptive generative data augmentation.

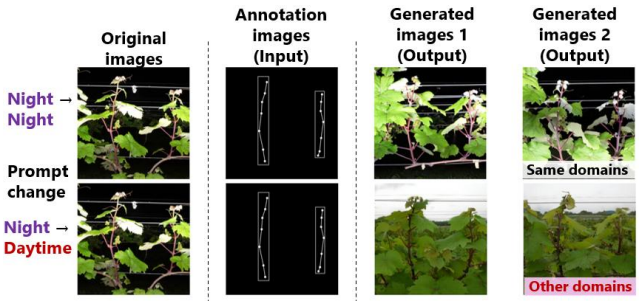

Figure 5: Generative data augmented images.

annotated images. The generated synthetic images adapt background textures, lighting conditions, and color distributions to the target domain while preserving the geometric structure of objects (stem contours and node positions). Since this conversion requires no annotated images from the target domain, D4 dramatically reduces annotation costs and enables cross-environmental generalization as shown in Figure 5.

To ensure data quality, D4 incorporates an auto-selection mechanism based on the DreamSim image quality metric, filtering generated images to guarantee structural similarity and realistic appearance. Only high-quality synthetic samples are included in the augmentation dataset, minimizing the risk of annotation inconsistencies or unrealistic images that could degrade performance.

Extensive experiments demonstrate that D4 is effective for both bounding box detection and keypoint detection in vine shoot phenotyping. Evaluations using YOLOv8, EfficientDet, DETR, HR-Net, and ViT-Pose showed D4 delivered consistent improvements across all metrics. Notably, with 1050 D4-generated samples, the YOLOv8 model achieved a +27.69% increase in mAP50 for branch bounding box detection and a +13.73% improvement in keypoint detection accuracy. These improvements clearly demonstrate that D4 successfully performed domain adaptation from night to daytime conditions while augmenting the training dataset with geometrically consistent synthetic images. Ablation studies further revealed that (1) Stage 1 pretraining is essential for stable domain transfer, (2) the quality of generated images strongly impacts detection performance, and (3) prompt design enhances the realism and structural consistency of generated samples.

By generating high-quality synthetic images, D4 effectively addresses several fundamental limitations of agricultural datasets: small sample sizes, the high cost and difficulty of manual annotation, and significant variability due to lighting, season, and crop condition. We believe this capability contributes to the development of more robust and versatile deep learning models for plant canopy phenotyping. Within our comprehensive Agri-DT framework, D4 is envisioned to serve as both a data generation module and a domain normalization mechanism. This will enhance model

Subsystem	Datasets	Environments	Model Architectures	Evaluation Metrics
Autonomous UGV for continuous monitoring	Unlabeled RGB-D videos collected via UGV patrols;	Commercial greenhouse;	**UGV-based data acquisition system** integrating depth sensing and fiducial-marker localization with autonomous docking and periodic RGB recording	Docking success rate: 100% (15/15); docking time: 18–35 s
Mobile Phenotyping of Grapevine reproductive organs	High-resolution vineyard RGB images (6000×4000)	Commercial vineyard (nighttime)	**Two-stage Mask R-CNN pipeline** for inflorescence and flower detection with patch-based processing and contrast enhancement	Inflorescence detection: F1=0.943; Individual flower detection: F1=0.903, R^2=0.981
D4: Diffusion-based domain-adaptive data augmentation	Unlabeled vineyard images + small annotated nighttime dataset (50 images)	Vineyard fields with domain variation (night and day)	**Text-guided diffusion model with ControlNet**, enabling annotation-preserving synthetic images via two-stage finetuning	Branch BBox detection: +27.69% mAP50; Keypoint detection: +13.73% accuracy.

Table 1: Summary of datasets, environments, model architectures, and evaluation metrics for each subsystem.

performance across seasons, cultivation methods, and sensor configurations while simultaneously reducing annotation efforts. Consequently, D4 supports scalable, crop-wide phenotypic analysis and contributes to building reliable, data-driven digital twins for agricultural systems. Although D4 was evaluated on grapevine shoots in this study, the annotation-conditional generation mechanism is not crop-specific. With appropriate bounding-box or keypoint annotations, the same pipeline can be applied to other plant organs while preserving geometric consistency.

Conclusion

This paper presents an integrated suite of technologies for constructing an extensible Agri-DT: (i) a low-cost UGV capable of autonomous charging and continuous greenhouse monitoring, (ii) a high-precision phenotypic analysis pipeline for detecting reproductive organs under variable field conditions, and (iii) a diffusion-based generative domain adaptation framework that enables data augmentation while preserving annotation consistency across heterogeneous environments. By synergistically combining these research outcomes—sustained robotic monitoring, fine-structure trait extraction, and domain-adaptive model training—we can form the data layer, perception layer, and modeling layer necessary for practical Agri-DT. This contributes to building an integrated pipeline capable of supporting growth simulation, anomaly detection, yield estimation, and crop behavior analysis. Table 1 summarizes the datasets, environments, model architectures, and evaluation metrics of each subsystem, providing an integrated view of the Agri-DT components. In addition, several limitations should be acknowledged. First, while the UGV achieved reliable long-term monitoring inside structured greenhouse lanes, navigation performance in more cluttered or dynamically changing environments remains to be validated. Second, although diffusion-based augmentation substantially reduces annotation effort, an initial amount of labeled data is still required to guide domain adaptation, and the cost–benefit ratio between expert annotation and synthetic generation warrants further study.

Future work will extend this foundation toward a more multimodal Agri-DT by incorporating additional sensing modalities (RGB-D, multispectral, thermal, physiological responses), cross-seasonal learning, and multi-view 3D canopy reconstruction. We also plan to explore real-time data assimilation, simulation-based hypothesis testing analysis, and closed-loop decision support using predictive control. By integrating field robotics, phenotyping, and generative AI, this research aims to realize a versatile and scalable Agri-DT framework that enhances production planning, sustainability, and resilience across diverse agricultural systems.

Acknowledgments

This research was partially supported by the Japan Science and Technology Agency (JST), FOREST Grant Number JPMJFR201B, Moonshot R&D Program (AI-ENGAGE), Grant Number JPMJMS25E6, and the Research Institute of Green Science and Technology Fund for Research Project Support (2025-RIGST-25101) of the National University Corporation, Shizuoka University, Japan.

References

Fasiolo, D.; Scalera, L.; Maset, E.; and Gasparetto, A. 2023. Towards Autonomous Mapping in Agriculture: A Review of Supportive Technologies for Ground Robotics. Computers and Electronics in Agriculture 169. doi: 10.1016/j.robot.2023.104514.

Hirahara, K.; Nakane, C.; Ebisawa, H.; Kuroda, T.; Iwaki, Y.; Utsumi, T.; Nomura, Y.; Koike, M.; and Mineno, H. 2025. D4: Text-Guided Diffusion Model-Based Domain Adaptive Data Augmentation for Vineyard Shoot Detection. *Computers and Electronics in Agriculture* 230. doi: 10.1016/j.compag.2024.109849 .

Muhammad, A.; Salman, Z.; Lee, K.; and Han, D. 2023. Harnessing the Power of Diffusion Models for Plant Disease Image Augmentation. *Frontiers in Plant Science.* 14:1280496. doi: 10.3389/fpls.2023.1280496.

Palacios, F.; Bueno, G.; Salido, J.; Diago, M.; Hernádez, I.; and Tardauila, J. 2020. Automated Grapevine Flower Detection and Quantification Method Based on Computer Vision and Deep Learning from On-the-go Imaging Using a Mobile Sensing Platform Under Field Conditions. *Computers and Electronics in Agriculture* 178. doi: 10.1016/j.compag.2020.105796.

Purcell, W., and Neubauer, T. 2023. Digital twins in agriculture: A state-of-the-art review. *Smart Agricultural Technology* 3: 100094. doi:10.1016/j.atech.2022.100094.

Rahim, U.; Utsumi, T.; and Mineno, H. 2022. Deep Learning-Based Accurate Grapevine Inflorescence and Flower Quantification in Unstructured Vineyard Images Acquired Using a Mobile Sensing Platform. *Computers and Electronics in Agriculture* 198: 107088. doi: 10.1016/j.compag.2022.107088.

Subeesh, A., and Chauhan, N. 2026. Agricultural digital twin for smart farming: A review. *Green Technologies and Sustainability* 4(2): 100299. doi:10.1016/j.grets.2025.100299.

Zobaer, A.; Watanabe, S.; Koike, M.; and Mineno, H. 2025. A Low-Cost Autonomous UGV with Self-Charging Capability for Efficient Greenhouse Monitoring. In *Proceedings of the IEEE Global Conference on Consumer Electronics (GCCE)*.