



SAPIENZA
UNIVERSITÀ DI ROMA

Overcoming the Annotation Bottleneck: Label-Efficient Deep Learning for Precision Agriculture and Remote Sensing

Department of Computer, Control and Management Engineering
Ph.D. in Engineering in Computer Science (XXXVIII cycle)

Ionut Marian Motoi

ID number 1892355

Advisor

Prof. Thomas Alessandro Ciarfuglia

Academic Year 2025/2026

Thesis defended on 21/05/2026

Overcoming the Annotation Bottleneck: Label-Efficient Deep Learning for Precision Agriculture and Remote Sensing

Ph.D. Thesis. Sapienza University of Rome

© 2026 Ionut Marian Motoi. All rights reserved

This thesis has been typeset by L^AT_EX and the Sapthesis class.

Author's email: motoi@diag.uniroma1.it

Abstract

The deployment of deep learning models in unstructured environments, such as precision agriculture and remote sensing, is frequently limited by the annotation bottleneck. Acquiring dense, high-quality labels in these domains is an expensive and time-consuming process, complicated by persistent covariate shifts, sparse targets, and pronounced class imbalances. This dissertation presents a cohesive methodological approach to overcome data scarcity by focusing on label-efficient learning strategies. The research utilizes two primary mechanisms to construct and amplify training signals: surrogate supervision, which derives structured targets from weak cues and structural priors, and compositional synthesis, which reshapes the effective training distribution by recombining object- or region-level instances.

The first part of the thesis focuses on precision agriculture, with a specific application to automated harvesting and monitoring in table grape vineyards. It introduces a weakly and semi-supervised pipeline that addresses domain shifts in object detection and instance segmentation through automated pseudo-label generation. To further reduce the reliance on manual field data collection, the research presents a hybrid data generation pipeline that blends real fruit instances into simulated 3D environments. This compositional synthesis approach is subsequently adapted for anomaly detection, utilizing foundational models and classical edge detection to generate realistic samples of rare agricultural defects. Additionally, the research addresses the data limitations associated with measuring internal fruit quality. It demonstrates that a multi-task deep neural network can accurately estimate Soluble Solid Content using low-cost RGB sensors, maintaining robust performance even when ground-truth labels are sparse, thereby providing a practical alternative to specialized instrumentation.

The second part transfers these core principles to the macroscopic scale of Earth observation. It adapts instance-level cut-and-paste augmentation to satellite semantic segmentation by extracting discrete objects from composite labels to improve model generalization. Building on this concept, the thesis introduces ChangeMix, a change-aware augmentation strategy for semi-supervised change detection. By maintaining a dynamic bank of high-confidence change regions and injecting them into unlabeled data while balancing temporal order, ChangeMix effectively manages the class imbalance and temporal bias characteristic of remote sensing datasets.

Across both domains, the experimental evaluations demonstrate that the proposed methodologies consistently improve model performance and robustness under limited labeling budgets. By maximizing the utility of available data and generating targeted synthetic supervision, this dissertation offers a scalable and accessible approach for deploying reliable perception systems in complex, real-world scenarios.

Acknowledgments

I would like to express my profound gratitude to my supervisor, Prof. Thomas Alessandro Ciarfuglia. His unwavering support and continuous encouragement throughout these years have not only guided this research but have also deeply shaped my growth as a researcher.

I would like to extend a special thanks to Prof. Daniele Nardi for his invaluable advice and mentorship. My gratitude also goes out to the other professors at the Department of Computer, Control and Management Engineering (DIAG) with whom I have had the privilege to collaborate. I am especially grateful to my colleagues Leonardo and Sandeep for all the wonderful moments we shared. To all the other colleagues, Ph.D. students, and researchers at the department, thank you for making the lab an enjoyable and lively place to work every day.

I want to warmly thank Prof. Xiaoxiang Zhu and the entire team at the Technical University of Munich (Chair of Data Science in Earth Observation) for welcoming me into their research group. My time there was highly formative, and I am grateful to my colleagues abroad for the fascinating discussions and great memories. A special thanks goes to Wei, Luigi, Constantin, Veronika, and Francesco, whose friendship made my experience in Munich truly unforgettable.

On a personal note, I want to thank my friends for always keeping me grounded and offering a much-needed escape from the world of research whenever it was needed. To my family, thank you for your unconditional love, your endless sacrifices, and for believing in me at every single step of my education. None of this would have been possible without you. Finally, to my girlfriend Elena, thank you for your immense patience and for standing by my side through all the highs and lows of this journey.

Contents

List of Figures	xi
List of Tables	xiii
1 Introduction	1
1.1 Context and Motivation	1
1.2 Research Objectives and Contributions	3
1.3 List of Publications	4
1.4 Thesis Outline	5
2 Background	7
2.1 Precision Agriculture	7
2.1.1 The Role of Perception and Robotics in Unstructured Environments	7
2.1.2 Case Study: The CANOPIES Project	7
2.1.3 Experimental Field	8
2.1.4 Tasks and Challenges	8
2.2 Remote Sensing	10
2.2.1 Characteristics of Satellite Imagery	10
2.2.2 Tasks and Challenges	11
2.3 Deep Learning Architectures	12
2.3.1 Feature Extractors	12
2.3.2 Object Detection Architectures	13
2.3.3 Dense Prediction Architectures	14
2.4 Learning Paradigms	15
2.4.1 Weakly-Supervised Learning	15
2.4.2 Semi-Supervised Learning	15
2.4.3 Foundation Models and Zero-Shot Transfer	16
2.5 Classical Computer Vision and Image Processing	16
2.5.1 Superpixel and Graph-based Segmentation	16
2.5.2 Edge Detection	17
I Precision Agriculture	19
3 Bridging Covariate Shift via Pseudo-Labeling	21
3.1 Materials and Methods	24

3.1.1	Target and Source Datasets	25
3.1.2	System Overview	26
3.1.3	Detection Pseudo-Label Generation	28
3.1.4	Segmentation Pseudo-Label Generation	29
3.1.5	Metrics	34
3.2	Results and Discussion	34
3.2.1	Object Detection Experiments	34
3.2.2	Instance Segmentation Experiments	36
3.2.3	Complete System Experiments	37
3.3	Concluding Remarks	38
4	Automated Hybrid Data Generation for Detection	41
4.1	Related Work	43
4.2	Materials and Methods	44
4.2.1	Synthetic Data Acquisition	44
4.2.2	Real-World Instance Extraction	45
4.2.3	Context-Aware Cut-and-Paste Blending	46
4.3	Experiments	47
4.3.1	Training Details	47
4.3.2	Evaluation Metrics	48
4.3.3	Results and discussion	49
4.4	Concluding Remarks	51
5	Synthetic Data Generation for Anomaly Detection	53
5.1	Materials and Methods	55
5.1.1	Data Collection and Labelling	55
5.1.2	Synthetic Data Generation Procedure	55
5.2	Experiments and Results	60
5.2.1	Baselines	60
5.2.2	Experiments	62
5.3	Concluding Remarks	63
6	Estimating SSC with Simple RGB Sensors	67
6.1	Hyperspectral and RGB Approaches	68
6.1.1	Hyperspectral applications to agricultural robotics	68
6.1.2	RGB applications to fruit quality assessment	70
6.2	Materials and Methods	71
6.2.1	Dataset	71
6.2.2	SSC Estimation Models	75
6.3	Experiments	80
6.3.1	Human Performance Baselines	80
6.3.2	Histogram-based Model Evaluation	82
6.3.3	Multi-task DNN Model Evaluation	82
6.4	Discussion	84
6.4.1	Summary of Performance and Comparison	85
6.4.2	Practical Implications and Trade-offs	86
6.4.3	Limitations and Future Work	86

6.5	Concluding Remarks	87
II	Remote Sensing	89
7	Cut-and-Paste Augmentation for Satellite Semantic Segmentation	91
7.1	Introduction	91
7.2	Methodology	92
7.2.1	Instance Extraction	92
7.2.2	Instance Pasting	94
7.3	Experiments	94
7.3.1	Dataset	94
7.3.2	Evaluation Metric	95
7.3.3	Training Details	95
7.3.4	Results	95
7.4	Concluding Remarks	96
8	ChangeMix: Change-Aware Cut-and-Paste Augmentation	99
8.1	Introduction	99
8.2	Related Work	102
8.2.1	Deep Learning-based Change Detection	102
8.2.2	Semi-Supervised Change Detection	102
8.2.3	Data Augmentation for Change Detection	103
8.3	Methodology	103
8.3.1	Problem Definition	103
8.3.2	Generalized Weak-to-Strong SSL Framework	104
8.3.3	ChangeMix: Change-Aware Cut-and-Paste Augmentation	105
8.4	Experiments	107
8.4.1	Datasets	107
8.4.2	Evaluation Metrics	108
8.4.3	Implementation Details	108
8.4.4	Comparison Experiments	109
8.4.5	Ablation Experiments	112
8.4.6	Hyperparameter Experiments	112
8.4.7	Qualitative Results	114
8.5	Concluding Remarks	116
9	Conclusion	119
9.1	Summary of Contributions	119
9.2	Broader Implications	120
9.3	Limitations and Future Directions	120
9.4	Concluding Remarks	121
	Bibliography	125

List of Figures

1.1	Methodological framework of the dissertation	2
2.1	(a) Satellite view of the experimental vineyard. (b) A picture of the trellis (Tendone) structure.	9
3.1	Representative examples of the four table grape varieties cultivated in the experimental vineyard	25
3.2	Visual comparison illustrating the covariate shift between the source and target datasets	26
3.3	Overall architecture of the proposed surrogate supervision framework	27
3.4	Workflow of the Detection Pseudo-Label Generation (DPLG) subsystem	28
3.5	Geometric verification process for robust instance association across consecutive video frames	29
3.6	The bounding box propagation mechanism	30
3.7	Workflow of the Segmentation Pseudo-Label Generation (SPLG) subsystem	30
3.8	Data flow within the Mask R-CNN architecture	31
3.9	Impact of external spatial priors on segmentation quality in the target domain	32
3.10	Examples of mask refinement strategies applied to initial pseudo-masks	33
3.11	Visual example of a pseudo-mask refined using the GrabCut algorithm	34
3.12	Qualitative comparison of object detection performance before and after domain adaptation	36
4.1	The target operational environment in Aprilia (Italy)	42
4.2	a) The simulated robotic platform, modeled after the Alitrak DCT-300P base and TIAGo++ manipulator, positioned in the virtual vineyard. (b) A synthetic image captured from the perspective of the arm-mounted camera.	43
4.3	Overview of the proposed hybrid data generation pipeline	45
4.4	Example of the automated mask extraction process	46
4.5	Example of a hybrid training image generated by the proposed pipeline	48
4.6	Qualitative comparison of YOLOv8 nano inference results across different test sequences	52
5.1	(a) Robotic harvesting platform and (b) localized anomalies	53
5.2	Architecture of the proposed compositional synthesis pipeline	56

5.3	Textural analysis of table grape patches using the Canny edge detector	57
5.4	Dual-Canny Edge Detection (DCED) filter applied to isolate surface defects	57
5.5	Geometric alignment and blending process for compositional synthesis	58
5.6	Qualitative examples of the compositional synthesis process	61
5.7	Qualitative classification results from the best-performing model on the validation set	64
6.1	The robotic harvesting platform deployed in the experimental vineyard	68
6.2	An overview of the experimental workflow	72
6.3	Examples of paired standard JPG and raw (white-balanced) images collected with the same smartphone device	72
6.4	Data collection featuring white balance correction and in-field berry sampling	74
6.5	Examples of the six color levels utilized for categorical quality classification	75
6.6	An illustration of the cross-shaped feature selection heuristic for histogram-based estimation	77
6.7	The multi-task network architecture utilized for joint SSC and color estimation	78
6.8	Correlation matrix of key parameters from the human evaluation study on 100 grape images	81
6.9	(a) Hyperparameter correlation with validation error (MAE). (b) Test error across varying λ and color correction.	83
6.10	Linear regression analysis of runs utilizing hyperparameters similar to the top 20 models on the <code>MixedSeason_{STD}</code> and <code>MixedSeason_{WB}</code> datasets	84
7.1	Overview of the proposed Cut-and-Paste augmentation technique . .	93
7.2	Qualitative visualization of semantic segmentation results on the internal validation set	97
8.1	Comparison of augmentation strategies and their impact on supervision	101
8.2	Overview of the proposed ChangeMix within a teacher-student semi-supervised change detection pipeline	106
8.3	Qualitative comparison of change detection results on the LEVIR-CD test set using a 10% label budget and the DPT architecture with a DINOv2-Base backbone	117
8.4	Qualitative comparison of change detection results on the WHU-CD test set using a 10% label budget and the DPT architecture with a DINOv2-Base backbone	118

List of Tables

3.1	Performance comparison between the source detector (SDet) and the target detector (TDet) on the test split of the static image target dataset (TImg)	35
3.2	Performance comparison between the source detector (SDet) and the target detector (TDet) on the test split of the video target dataset (TVid)	35
3.3	Quantitative evaluation of the covariate shift impact on the baseline Mask R-CNN model (SSeg), trained exclusively on the source dataset (WGISD)	37
3.4	Comparison of pseudo-mask refinement strategies	37
3.5	End-to-end system evaluation on the TImg test set	38
4.1	Quantitative performance evaluation of the YOLOv8-nano model across four distinct test sequences	50
4.2	Quantitative performance evaluation of the YOLOv8-small model across four distinct test sequences	51
5.1	Performance comparison of the baseline models	62
5.2	Performance impact of augmenting or substituting the training set with single-berry synthetic anomaly samples	63
5.3	Performance impact of augmenting or substituting the training set with multi-berry (three instances) synthetic anomaly samples	64
6.1	Comparison of Hyperspectral (HSI) and RGB Imaging devices for in-field use	69
6.2	Description of the atomic splits according to the device, image format, gray card presence, and the possibility of having expert ground truth from white-balanced images	73
6.3	Summary of the experimental dataset splits used for model training and evaluation	74
6.4	Statistical properties of the experimental dataset splits	76
6.5	Master comparison of model performance against baselines and state-of-the-art methods	85
7.1	Validation performance of the Cut-and-Paste augmentation compared to the baseline U-Net model	96

7.2	Test set performance evaluating the optimal Cut-and-Paste augmentation configurations ($N = 100$) against the baseline model	98
8.1	Comparison of semi-supervised change detection methods on the LEVIR-CD test set	110
8.2	Comparison of semi-supervised change detection methods on the WHU-CD test set	111
8.3	Backbone comparison for UniMatch v2 and the ChangeMix augmentation on the LEVIR-CD test set	112
8.4	Backbone comparison for UniMatch v2 and the ChangeMix augmentation on the WHU-CD test set	113
8.5	Ablation study on the components of ChangeMix	113
8.6	Instance Bank hyperparameter evaluation on LEVIR-CD (20% labels) using the DPT architecture with DINOv2-Base and DINOv2-Small .	115
8.7	Instance Bank hyperparameter evaluation on WHU-CD (20% labels) using the DPT architecture with DINOv2-Base and DINOv2-Small .	116

Chapter 1

Introduction

1.1 Context and Motivation

The advent of Deep Learning has fundamentally transformed the field of Computer Vision, enabling machines to perceive, classify, and segment the visual world with remarkable accuracy [76, 80]. This progress has been largely driven by the availability of massive, curated datasets such as ImageNet [32] or Microsoft COCO [89], where the supervisory signal is abundant and the relevant visual patterns are well represented. When such algorithms transition from controlled digital benchmarks to physical environments, the assumption of high-quality data availability frequently no longer holds. This transition often introduces covariate shift, where the data distribution encountered during inference differs significantly from the training distribution [138]. In unstructured environments, such distributional shifts are the norm, driven by changing seasons, varying sensors, and diverse geographies [118]. Many practical perception tasks compound this challenge with a second form of data inefficiency in which the informative signal is limited, localized, or weakly observable. Consequently, the target signal can be overwhelmed by dominant background context.

In Precision Agriculture [167], for example in collaborative robotic harvesting for table grape vineyards [1], the informative signal is concentrated in discrete, partially occluded instances such as grape bunches or individual berries, while the surrounding canopy constitutes the vast majority of the visual scene. Data collection is seasonal, labor-intensive, and subject to pronounced environmental variability [34]. Moreover, some target properties are not directly observable from images. Ground truth for internal quality attributes such as fruit sugar content typically requires specialized instruments, such as contact spectrometers or analog refractometers, rather than visual inspection [68, 141]. As a result, both the acquisition of representative images and the collection of reliable labels become tightly constrained resources.

Similarly, in Remote Sensing, the vast geographic diversity of the Earth’s surface induces substantial covariate shift. Furthermore, many high-impact tasks, particularly change detection, often exhibit the same scarcity pattern where the event of interest (e.g., a new building) is spatially sparse and statistically rare, while stable background context dominates the image. In major benchmark datasets, the changed class can occupy less than 5% of the pixels, making standard supervised learning inefficient and sensitive to dataset bias [64, 87, 114].

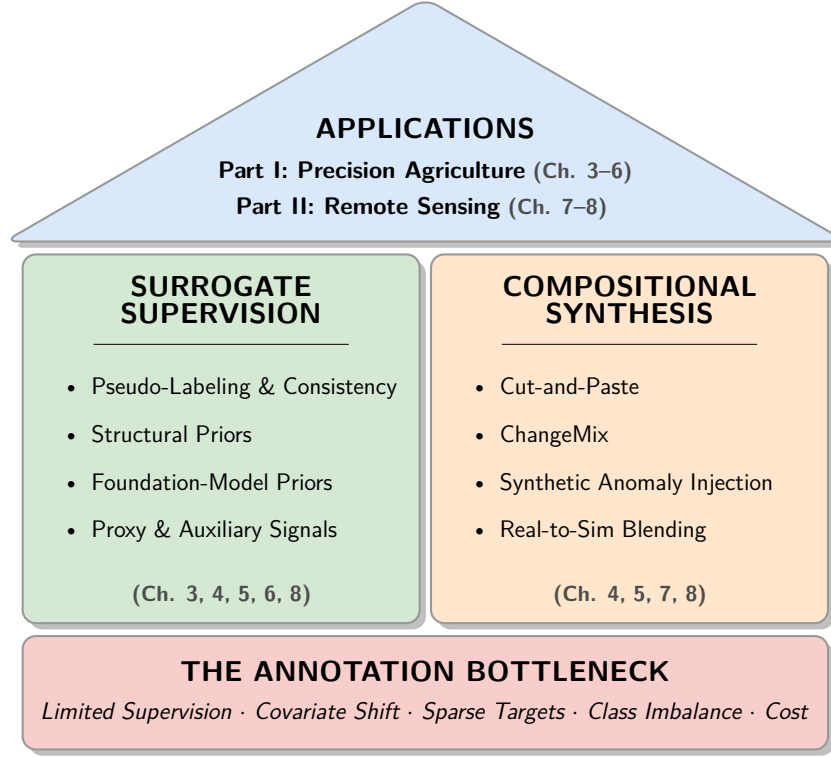


Figure 1.1. Methodological framework of the dissertation. The central challenge of the Annotation Bottleneck is addressed via two core strategies, Surrogate Supervision and Compositional Synthesis, which enable robust perception applications in Precision Agriculture and Remote Sensing.

Across both domains, performance is typically constrained more by the availability, coverage, and quality of supervision than by model capacity. Consequently, the primary obstacle is the annotation bottleneck. Acquiring reliable training supervision is costly and difficult to scale, whether the task requires dense segmentation masks, temporal changes, anomaly labels, or instrument-based measurements of internal properties [84, 150]. This bottleneck is further amplified by domain shift, class imbalance, and temporal bias, which reduce the effectiveness and transferability of small labeled sets.

To address this, the dissertation adopts a data-centric strategy that focuses on how supervision can be constructed and amplified under limited labeling budgets and imperfect annotations. The approach relies on two recurring mechanisms. Surrogate supervision derives structured training targets from weak cues, structural constraints, or foundation-model priors. Compositional synthesis increases the frequency and diversity of informative regions by extracting and recombining object-level and region-level instances, reshaping the effective training distribution.

As illustrated in Figure 1.1, the dissertation is organized into two parts, connected by the shared methodological thread introduced above. Part I (Precision Agriculture) begins in table grape vineyards, where supervision is constrained by seasonal acquisition, strong environmental variability, and the difficulty of measuring certain targets. Surrogate supervision is constructed via weakly and semi-supervised

learning, generating pseudo-labels from labeled source data and unlabeled target-domain video to reduce reliance on dense manual annotation and mitigate covariate shift [26]. Foundation models were also leveraged to extract instance-level examples of interest, including both real fruit instances and candidate anomalous regions [129, 99]. Compositional synthesis is then pursued through synthetic environments and instance-level mixing to expand the diversity of training signals under small datasets [129, 99]. Finally, the feasibility of using low-cost RGB sensing to support the estimation of internal quality attributes, such as sugar content, is investigated, reducing dependence on specialized hardware [27].

Part II (Remote Sensing) transfers the same methodological framework to Earth Observation, where annotation constraints persist and are compounded by extreme class imbalance and temporal bias. Instance injection and Cut-and-Paste style augmentation are adapted to satellite imagery to reshape the effective training distribution and improve generalization in segmentation tasks [101]. In semi-supervised change detection, ChangeMix [100] is introduced, which accumulates high-confidence change regions as training targets and injects them into unlabeled data, while randomly swapping the temporal order to model both appearance and disappearance. This mitigates temporal bias and improves learning under severe class imbalance and limited label budgets.

Across both domains, the thesis shows that robust performance under limited and imperfect supervision can be achieved by employing surrogate supervision and compositional synthesis across multiple computer vision tasks.

1.2 Research Objectives and Contributions

The primary objective of this thesis is to develop label-efficient Deep Learning methodologies that enable robust perception in unstructured environments where supervision is scarce, imbalanced, noisy, expensive to acquire, or difficult to standardize across operating conditions. Accordingly, this research investigates how to construct and amplify supervision so that models remain reliable across varying operational scenarios. The resulting methodology is built around two recurring mechanisms introduced in Section 1.1: surrogate supervision, which derives structured training signals from weak cues and priors, and compositional synthesis, which reshapes the effective training distribution by recombining object-level or region-level instances. These mechanisms are validated across two application domains, Precision Agriculture and Remote Sensing, demonstrating transferability across markedly different sensing scales and deployment constraints.

The specific contributions of this dissertation are organized into two parts.

Part I: Precision Agriculture

- **Bridging Covariate Shift via Pseudo-Labeling (Chapter 3):** A weakly and semi-supervised system is proposed to mitigate the domain gap between available source datasets and specific target fields. By leveraging bounding box priors and refinement strategies built upon classical computer vision techniques, a pipeline is presented that generates high-quality segmentation pseudo-labels

from simple smartphone videos, significantly reducing the manual annotation effort required for domain adaptation.

- **Automated Hybrid Data Generation for Detection (Chapter 4):** A data generation pipeline is introduced that leverages simulated environments to augment limited real-world data. By automatically extracting masks of actual grape instances and integrating them into diverse synthetic vineyard scenes with geometric consistency, realistic training samples are generated. This approach improves detection performance without requiring large-scale field collection.
- **Synthetic Data Generation for Anomaly Detection (Chapter 5):** To address the limited availability of labeled anomalies, such as defects, diseases, and pests, a method is developed that combines foundation-model segmentation priors with texture-aware cues to synthesize realistic anomalous samples. By isolating anomalous textures and seamlessly blending them onto healthy fruit, this approach improves anomaly recognition in low-supervision regimes.
- **Estimating SSC with Simple RGB Sensors (Chapter 6):** Moving beyond geometric perception, a methodology is proposed to infer internal fruit quality, specifically Soluble Solid Content (SSC), using low-cost RGB sensors. A multi-task Deep Neural Network is introduced that jointly estimates agronomic color class and sugar content, offering a cost-effective, non-destructive alternative to specialized instrumentation.

Part II: Remote Sensing

- **Cut-and-Paste Augmentation for Satellite Semantic Segmentation (Chapter 7):** Instance-level augmentation is adapted to semantic segmentation in satellite imagery. By extracting discrete connected components from composite semantic labels to create pasteable instances, Cut-and-Paste style augmentation is shown to improve generalization and segmentation performance in Land Use Land Cover (LULC) mapping.
- **ChangeMix: Change-Aware Cut-and-Paste Augmentation (Chapter 8):** ChangeMix is introduced as an augmentation strategy for semi-supervised change detection designed to address severe class imbalance and temporal bias. The method maintains a dynamic instance bank of high-confidence change regions and injects them into unlabeled data, while randomly swapping the temporal order to model both appearance and disappearance events. The evaluation demonstrates consistent performance improvements under limited label budgets across multiple architectures and frameworks.

1.3 List of Publications

1. Thomas A. Ciarfuglia¹, Ionut M. Motoi¹, Leonardo Saraceni¹, Mulham Fawakherji, Alberto Sanfeliu, and Daniele Nardi. "Weakly and semi-supervised detection, segmentation and tracking of table grapes with limited and noisy data." *Computers and Electronics in Agriculture* 205 (2023): 107624 [26]

2. **Ionut M. Motoi**¹, Valerio Belli¹, Alberto Carpineto¹, Daniele Nardi, and Thomas A. Ciarfuglia¹. "Synthetic data generation for anomaly detection on table grapes." *Smart Agricultural Technology* 10 (2025): 100787. [99]
3. Leonardo Saraceni, **Ionut M. Motoi**, Daniele Nardi, and Thomas A. Ciarfuglia. "Self-supervised data generation for precision agriculture: Blending simulated environments with real imagery." In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pp. 71-77. IEEE, 2024. [129]
4. Thomas A. Ciarfuglia, **Ionut M. Motoi**, Leonardo Saraceni, and Daniele Nardi. "Can robots "Taste" grapes? Estimating SSC with simple RGB sensors." *Computers and Electronics in Agriculture* 239 (2025): 110845. [27]
5. **Ionut M. Motoi**, Leonardo Saraceni, Daniele Nardi, and Thomas A. Ciarfuglia. "Evaluating the efficacy of cut-and-paste data augmentation in semantic segmentation for satellite imagery." In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 9802-9806. IEEE, 2024. [101]
6. **Ionut M. Motoi**, Wei Huang, Thomas A. Ciarfuglia, Daniele Nardi, and Xiao Xiang Zhu. "ChangeMix: Change-Aware Cut-and-Paste Augmentation for Semi-Supervised Remote Sensing Change Detection." (submitted, under review). [100]
7. Leonardo Saraceni, **Ionut M. Motoi**, Daniele Nardi, and Thomas A. Ciarfuglia. "Agrisort: A simple online real-time tracking-by-detection framework for robotics in precision agriculture." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2675-2682. IEEE, 2024. [128]
8. Marc Dalmaso, Francisco Garcia-Ruiz, Thomas A. Ciarfuglia, Leonardo Saraceni, **Ionut M. Motoi**, Emilio Gil, and Alberto Sanfeliu. "Human-robot harvesting plan negotiation: perception, grape mapping and shared planning." In *2024 7th Iberian Robotics Conference (ROBOT)*, pp. 1-8. IEEE, 2024. [30]

1.4 Thesis Outline

The thesis is organized as follows:

- Chapter 1 presents the research context, objectives, and main contributions.
- Chapter 2 introduces the two application domains and their challenges, and summarizes key technical foundations used throughout the thesis.
- Chapters 3 through 8 present the contributions developed during the research period.
- Chapter 9 concludes with a summary of the findings and a discussion of future research directions.

¹The authors contributed equally.

Chapter 2

Background

2.1 Precision Agriculture

The evolution of modern farming is frequently categorized into stages, with the current paradigm often termed Agriculture 4.0 [167]. This stage is characterized by the integration of the Internet of Things (IoT), Big Data, Artificial Intelligence (AI), and Robotics. The primary objective of these technologies is to increase the quantity and quality of agricultural products while optimizing the use of inputs, such as water, energy, fertilizers, and pesticides, thereby reducing the environmental footprint and operational costs.

2.1.1 The Role of Perception and Robotics in Unstructured Environments

Agricultural robotics has emerged as a prominent technology within the Precision Agriculture framework to address pressing issues such as labor shortages and the need for improved operational efficiency [105]. However, deploying robots in agricultural settings presents challenges significantly different from those found in industrial automation. Industrial environments are typically structured, predictable, and controlled. In contrast, agricultural environments are unstructured, dynamic, and subject to substantial variability in illumination, weather conditions, and object appearance [75, 46].

For a robot to interact effectively with a crop, whether for harvesting, pruning, or spraying, it must possess robust perception capabilities. It must detect fruit amidst dense foliage, estimate ripeness from visual cues, and operate reliably despite the covariate shifts caused by changing seasons or different crop varieties [5, 105]. This thesis addresses these specific perception bottlenecks, focusing on data-efficient deep learning methods to bridge the gap between controlled experimental settings and the complex, dynamic realities encountered in real-world field environments.

2.1.2 Case Study: The CANOPIES Project

The research presented in Part I of this thesis is grounded in the challenges and requirements of the CANOPIES project¹ [1] (A Collaborative Paradigm for Human

¹<https://canopies.inf.uniroma3.it/>

Workers and Multi-Robot Teams in Precision Agriculture Systems). Funded under the European Union’s Horizon 2020 research and innovation program ², CANOPIES represents a significant shift from fully autonomous industrial automation toward a collaborative human-robot paradigm in unstructured agricultural environments.

The project focuses on permanent crops, and in particular on table grape vineyards. By introducing teams of robots that work alongside human farmers, it addresses the challenges of labor shortages and efficiency. The system employs a heterogeneous team of robots to perform agronomic tasks such as harvesting and pruning:

- **Farming Robot:** A mobile manipulator equipped with a bimanual torso, designed to perform delicate interactions with the crop, such as detecting grape bunches, assessing their ripeness, and finally harvesting them.
- **Logistics Robot:** An autonomous mobile platform responsible for transporting harvested crates.

While the project’s broader scope includes robotics, human-robot interaction, and logistics, it serves in this thesis as the primary domain for applying and validating Deep Learning and Computer Vision methodologies in Precision Agriculture. The research contributions were developed prioritizing computational constraints, ensuring models are lightweight enough to execute on mobile edge platforms while reducing the reliance on labeled data. This approach also aims to democratize their utility beyond specialized robotic systems to more accessible, cost-effective equipment.

2.1.3 Experimental Field

The experimental activities [26] were conducted in a commercial vineyard located in Aprilia, southern Lazio (Italy). The site comprises two plots totaling approximately 1.16 hectares (Figure 2.1a). The vineyards are structured as a traditional trellis system known as *Tendone* (Figure 2.1b), with a wide 3×3 meters spacing between plants. The vines are mature (older than 3 years) and in full production, representing typical working conditions for agronomic tasks such as fruit harvesting or vine pruning. The structures are covered with plastic sheeting and protective nets to shield grapes from rain and hail.

2.1.4 Tasks and Challenges

The deployment of a robotic platform in a vineyard requires a hierarchy of perceptual capabilities. The robot must first locate the fruit within the canopy, understand its precise shape for manipulation, verify its health, and finally assess its ripeness to make a harvesting decision. The choice of table grapes as the target crop introduces specific computer vision challenges that differ significantly from those encountered with other fruits or in structured industrial settings.

²Grant Agreement No. 101016906



Figure 2.1. (a) Satellite view of the experimental vineyard. The image shows the varieties grown in each row. (b) A picture of the trellis (Tendone) structure.

Detection and Instance Segmentation

While object detection is often sufficient for tasks such as yield estimation or simple counting, robotic manipulation requires a finer understanding of the object's geometry. Table grapes present a unique challenge compared to rigid fruits like apples or oranges; they form complex, deformable clusters where individual berries are often self-similar and heavily occluded by foliage or other bunches [166, 127]. Furthermore, the visual appearance of the bunch varies drastically depending on the viewing angle and lighting conditions, which can shift rapidly in an outdoor environment. Unlike wine grapes, which can be mechanically harvested using straddle tractors, table grapes require delicate handling to prevent berry damage, which would compromise their marketability and result in economic losses [170, 65]. To enable a robotic hand to approach and grasp the bunch stem without damaging the berries, the vision system must provide precise instance segmentation masks. This requires algorithms capable of delineating the exact boundaries of the fruit mass, separating it from the background and adjacent clusters, a task that demands high-quality pixel-level annotations which are expensive to acquire.

Anomaly Detection

Beyond locating the fruit, the system must identify potential defects such as visible mold, rot, or pest infestations [86]. Early detection is essential to prevent the spread of pathogens and to ensure that only healthy fruit is harvested or stored. The primary challenge in training models for this task is the scarcity of data [97]. Anomalies are, by definition, rare events. Finding enough naturally occurring examples to train a robust supervised classifier is often impractical. This creates a strong motivation for synthetic data generation methods capable of simulating these rare textural defects.

Quality and Ripeness Estimation

Finally, an autonomous harvesting robot must replicate the decision-making process of a human expert regarding fruit ripeness. For table grapes, this is largely determined by the Soluble Solids Content (SSC), commonly measured in °Brix. Standard practice

relies on refractometers [141], which require destructive sampling, or contact-based spectrometers [68] to measure SSC. While effective for spot checks, these manual methods create a significant bottleneck for data-driven approaches. Unlike tasks such as object detection, where visual inspection of images suffices for labeling, acquiring ground truth for internal fruit quality is inherently slow, invasive, and often destructive. This difficulty results in sparse datasets that make training standard deep learning models problematic. Conversely, a robotic solution ideally requires non-invasive, remote estimation. The challenge, therefore, lies in inferring this internal chemical property from external visual cues [12, 158, 18], such as color and skin texture, using standard RGB sensors. This relationship is often latent and influenced by environmental factors [62], necessitating robust label-efficient modeling strategies to overcome the scarcity of ground truth data.

The Data Scarcity Bottleneck.

Across all these tasks, the unifying bottleneck faced when deploying Deep Learning models in real-world scenarios, such as the CANOPIES project, is the scarcity of annotated data [70, 75]. Collecting and labeling high-resolution agricultural datasets is labor-intensive, seasonal, costly, and in problems like anomaly detection, requires expert annotation. This fundamental limitation provides the primary motivation for the data-efficient methodologies developed in Part I.

2.2 Remote Sensing

While Part I of this thesis addresses perception in the proximal environment of the vineyard, Part II shifts the perspective to the global scale of Earth Observation. Remote Sensing plays an essential role in monitoring the Earth’s surface, driving several applications for sustainable development, such as urban planning, disaster response, and environmental conservation [4]. These applications directly support the United Nations Sustainable Development Goals (SDGs), particularly in building sustainable cities and taking climate action [72].

Much like Precision Agriculture, the Remote Sensing domain is characterized by distinct data inefficiencies. The fundamental barrier to deploying robust deep learning models remains the scarcity of high-quality, task-specific annotations [95, 174].

2.2.1 Characteristics of Satellite Imagery

Satellite imagery differs fundamentally from the proximal images captured by agricultural robots. It is characterized by varying acquisition sensors, each with different spatial, spectral, and temporal resolutions. Public initiatives, such as the European Space Agency’s Copernicus program, provide open-source multispectral imagery via Sentinel-2 with a revisit time of about five days [35]. However, its spatial resolution with a Ground Sample Distance (GSD) of 10 to 60 meters per pixel is often insufficient for fine-grained tasks.

High-resolution applications often utilize commercial data, such as PlanetScope imagery (approximately 3 meters per pixel). However, precisely detecting structural changes in buildings demands Very High Resolution (VHR) imagery with a GSD of

less than one meter [43]. Although this level of detail is a prerogative for granular structural analysis, it involves significantly higher acquisition costs.

The commercial burden, combined with the expensive expert labor required for pixel-perfect annotation [28, 153], poses a substantial constraint for Small and Medium Enterprises (SMEs) and academic researchers [79]. These practitioners must often rely on limited public datasets, requiring algorithms capable of maximizing performance from scarce data [153].

Moreover, from a computer vision perspective, these images introduce specific challenges. Unlike the object-centric view in agriculture, where a camera focuses on single fruits, satellite imagery is scene-centric [23]. A single image may cover kilometers of diverse terrain, introducing massive intra-class variability due to geographic diversity, seasonal changes, and atmospheric conditions [23].

2.2.2 Tasks and Challenges

This research focuses on two primary tasks in remote sensing.

Semantic Segmentation

Semantic segmentation in satellite imagery involves categorizing every pixel into classes such as impervious surfaces, forests, water bodies, or agricultural land. This process, often referred to as Land Use Land Cover (LULC) classification, is essential for interpreting surface dynamics [135]. The primary challenge here mirrors the segmentation problem explored in precision agriculture: generating pixel-perfect annotations for vast geographic areas is prohibitively expensive and labor-intensive [95, 140]. Furthermore, the resulting datasets often suffer from pronounced class imbalance, where dominant background classes like water or vegetation can overshadow rarer but significant classes like urban settlements, increasing the difficulty for models to accurately segment the minority classes [135, 101].

Change Detection

Change Detection (CD) is the task of identifying transitions in the state of a geographic area by comparing co-registered observations acquired at different timestamps [64]. Several high-impact applications build upon this task, including urban sprawl monitoring or post-disaster damage assessment [150, 31].

Similar to LULC classification, CD is significantly constrained by an annotation bottleneck and intrinsic class imbalance. However, these challenges are often more pronounced in CD. When analyzing temporal pairs of satellite imagery, the vast majority of pixels (frequently exceeding 95%) represent semantic stability (no change) [87, 57, 114]. This skew often leads standard deep learning models to develop a systematic bias toward the majority "no change" class, resulting in high global accuracy but poor recall for actual change events [83].

Moreover, real-world datasets often exhibit a temporal asymmetry between change types, frequently favoring object appearance (e.g., building construction) over object disappearance (e.g., demolition). This bias can lead to models that fail to detect significant events simply because they follow a temporal trajectory that is underrepresented in the training data [131, 100].

These challenges confirm that across both Precision Agriculture and Remote Sensing, the primary obstacle is not a scarcity of raw data, but rather a lack of high-quality, curated labels that adequately represent the minority classes of interest.

Therefore, Part II of this thesis extends the Cut-and-Paste and Semi-Supervised paradigms developed in the vineyard to the satellite domain. The evaluation demonstrates that the core principle of data augmentation through instance injection remains a robust strategy. By adapting these implementations to the domain-specific constraints, the proposed methodologies effectively mitigate the class imbalance and label scarcity issues inherent to both fields.

2.3 Deep Learning Architectures

The transition from manual feature engineering to end-to-end Deep Learning has significantly advanced computer vision [80, 76]. This dissertation utilizes several state-of-the-art architectures, ranging from efficient one-stage detectors suitable for real-time robotic applications to complex transformer-based models capable of capturing long-range dependencies in satellite imagery. This section outlines the core principles of these architectures, categorized by their primary role in this research: feature extraction, object detection, and dense prediction.

2.3.1 Feature Extractors

The feature extraction networks, commonly referred to as backbones, are the foundation of modern vision systems. They are responsible for projecting raw pixel data into a compressed, high-level feature representation space.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) represent a fundamental methodology in modern deep learning for computer vision [76]. Unlike traditional algorithms that relied on handcrafted features, CNNs learn to extract hierarchical visual patterns directly from raw pixels [81]. The core mechanism involves sliding learnable filters across the input to produce feature maps. The networks are structured hierarchically, with the initial layers typically detecting low-level features such as edges and texture, while deeper layers aggregate these patterns to recognize complex, high-level semantic concepts. This abstraction is achieved by alternating convolutional layers, non-linear activation, and pooling operations, which progressively reduce spatial resolution while increasing the depth of the feature representation. CNNs explicitly leverage the spatial locality and translational invariance inherent in natural images.

Residual Networks

Deep CNNs often suffer from the vanishing gradient problem, leading to performance saturation or degradation as network depth increases. ResNet addresses this issue by introducing residual learning, which utilizes skip connections to allow gradients to flow more easily during backpropagation [56]. Instead of learning a direct mapping $H(x)$, the network learns a residual function $F(x) = H(x) - x$, such that the output

is $y = F(x) + x$. This architecture serves as a robust and efficient feature extractor for several contributions in this thesis, including the instance segmentation of table grapes in Chapter 3, the multi-task auto-encoder for sugar content estimation in Chapter 6, and the change detection baselines in Chapter 8.

Vision Transformers

While CNNs process images through local receptive fields, Vision Transformers (ViTs) treat images as sequences of patches, outputting a sequence of tokens rather than grid-like feature maps [33]. By leveraging self-attention mechanisms, ViTs capture global dependencies, allowing the model to attend to different parts of an image simultaneously regardless of their spatial distance. In this thesis, DINOv2 [106] backbones are utilized, which are ViTs pre-trained on large unlabeled datasets via self-supervised learning. This pre-training strategy aligns directly with the label-efficient objectives of this research. By initializing models with robust, semantic features learned without manual supervision, the need to learn low-level visual features from scratch is effectively bypassed, making this approach particularly effective for downstream tasks where annotations are scarce. DINOv2 is available in different sizes. Specifically, the DINOv2-Small and DINOv2-Base variants serve as the backbones for the semi-supervised change detection architecture in Chapter 8.

2.3.2 Object Detection Architectures

Object detection goes beyond simple image classification by simultaneously localizing and classifying multiple objects within an image using bounding boxes. Modern deep learning approaches to this task are broadly categorized into two paradigms: two-stage and one-stage detectors [91]. Two-stage detectors, such as Faster R-CNN [123], first generate a set of candidate region proposals and then classify and refine them. While they achieve high accuracy, their sequential nature makes them computationally intensive. In contrast, for applications in agricultural robotics where real-time performance on edge hardware is critical, this thesis prioritizes one-stage detectors.

YOLO (You Only Look Once)

Unlike region-based approaches, YOLO [121] frames object detection as a single regression problem. It divides the input image into an $S \times S$ grid, where each cell is responsible for predicting bounding box coordinates and class probabilities simultaneously. This single-stage architecture balances inference speed and detection accuracy, making it highly suitable for onboard robotic perception. Furthermore, the architectural simplicity of single-stage detectors often makes them more robust when trained with noisy or imperfect supervision compared to complex two-stage models. Since its introduction, the YOLO architecture has undergone numerous iterations by different authors, progressively improving its feature extraction and anchor mechanisms. In this dissertation, YOLOv5 [66] is utilized as the core detector for the pseudo-label generation pipeline in Chapter 3 and YOLOv8 [67] is employed for the synthetic data generation pipeline in Chapter 4.

2.3.3 Dense Prediction Architectures

Dense prediction tasks, such as semantic and instance segmentation, extend beyond bounding box localization by assigning a class label to every individual pixel (semantic segmentation) or delineating the exact boundaries of distinct object instances (instance segmentation).

Mask R-CNN

Mask R-CNN [55] is a two-stage architecture for instance segmentation that extends the Faster R-CNN [123] object detector by adding a parallel branch for predicting segmentation masks. Faster R-CNN introduced the Region Proposal Network (RPN) for proposing candidate regions of interest. Mask R-CNN builds upon this by introducing the RoIAlign layer, which utilizes bilinear interpolation to extract features from each Region of Interest (RoI) with sub-pixel alignment, correcting the spatial quantization errors of previous methods and enabling precise mask generation. Mask R-CNN is employed in Chapter 3 for the instance segmentation of table grapes. Specifically, its standard inference flow is modified to accept external bounding box priors for pseudo-label generation.

U-Net

Originally designed for biomedical image segmentation, U-Net [124] is characterized by its symmetric Encoder-Decoder structure. The encoder captures the broad semantic context through progressive downsampling, while the decoder recovers precise spatial localization through upsampling. Fundamental to its success are skip connections that concatenate high-resolution features from the encoder directly to the decoder, preserving spatial details typically lost during pooling operations. Despite the emergence of numerous newer architectures, U-Net remains highly relevant due to its broad adoption, high applicability to different domains, and competitive performance, often comparable to other State of the Art (SotA) architectures. This architecture serves as the auto-encoder backbone for the multi-task sugar content estimation network in Chapter 6, and it acts as the baseline model for satellite semantic segmentation in Chapter 7.

DeepLabV3+

For semantic segmentation in remote sensing, capturing context at multiple scales can prove valuable. DeepLabV3+ [21] utilizes Atrous (Dilated) Convolutions to expand the receptive field without sacrificing spatial resolution. In particular, it employs an Atrous Spatial Pyramid Pooling module to probe features at multiple effective fields of view simultaneously. DeepLabV3+ is utilized in Chapter 8 to validate the model-agnostic capabilities of the ChangeMix augmentation strategy.

Dense Prediction Transformers

Dense Prediction Transformers (DPT) [120] is an architecture designed to leverage Vision Transformers for dense prediction tasks such as segmentation and depth

estimation. It reassembles the tokens from various stages of the ViT backbone into image-like feature representations, which are then progressively fused in a convolutional decoder. This architecture is extensively used in Chapter 8 for change detection.

2.4 Learning Paradigms

The central challenge addressed in this thesis is the scarcity of high-quality, task-specific annotations in the domains of Precision Agriculture and Remote Sensing. Standard Supervised Learning, which requires a manual label for every training sample, is often prohibitively expensive or impractical in these contexts. To overcome this annotation bottleneck, this research relies on alternative learning paradigms that maximize performance from limited or noisy supervision.

2.4.1 Weakly-Supervised Learning

Weakly-Supervised Learning (WSL) involves training models using "weak" annotations [172]. These labels are usually cheaper and faster to acquire than those required for full supervision, though they are coarser and contain less information. For example, in Chapter 3, the challenge of training instance segmentation models is addressed by using only bounding box annotations on the target domain, rather than standard pixel-perfect masks [26, 73]. By leveraging the spatial priors provided by these bounding boxes, the model can infer the correct number of objects and the underlying shapes, significantly reducing the annotation effort required for complex agricultural scenes.

2.4.2 Semi-Supervised Learning

Semi-Supervised Learning (SSL) leverages a small set of labeled data in conjunction with a large pool of unlabeled data [145]. The core assumption is that the structure of the unlabeled data distribution can provide valuable learning signals to regularize the model and improve generalization. This thesis employs two primary SSL mechanisms:

Pseudo-Labeling

Also known as self-training, this method involves using a model trained on labeled data to generate predictions (pseudo-labels) for unlabeled samples [82]. High-confidence predictions are then treated as ground truth to train the model. In Chapter 3, pseudo-labeling is utilized to adapt a segmentation model from a source domain to a target vineyard environment. Similarly, in Chapter 8, pseudo-labels are employed to populate a dynamic "Instance Bank" of change detection examples from unlabeled satellite imagery.

Consistency Regularization

Consistency regularization enforces the principle that a model's predictions should remain stable when the input is perturbed, provided the semantic content remains unchanged. In a typical setup (e.g., FixMatch [134] or UniMatch [162]), the model

is presented with a "weakly augmented" view and a "strongly augmented" view of the same unlabeled image. The prediction on the weak view serves as the supervision target for the strong view. Theoretically, this enforces the low-density separation principle, encouraging the decision boundary to pass through low-density regions of the data distribution [48]. However, in scenarios characterized by extreme class imbalance, such as Change Detection, standard random perturbations (e.g., color jitter, blur) are statistically unlikely to affect the sparse minority class pixels. Consequently, the consistency loss is dominated by the majority "background" class signal. In Chapter 8, this limitation is addressed by introducing targeted, class-aware perturbations via the ChangeMix strategy. This ensures that the consistency mechanism is actively applied to the rare features of interest, maximizing the supervision signal derived from the unlabeled data.

2.4.3 Foundation Models and Zero-Shot Transfer

The recent emergence of Foundation Models, which are large-scale models trained on vast and diverse datasets, has introduced a new paradigm for data efficiency [15]. These models learn robust, general-purpose representations that can be transferred to downstream tasks with little to no fine-tuning. Specifically, this thesis leverages the Segment Anything Model (SAM) [74]. SAM is a promptable segmentation architecture capable of generating high-quality masks from simple inputs such as points or bounding boxes. In Chapters 4 and 5, SAM's zero-shot capabilities are exploited to automatically extract object masks for synthetic data generation, effectively bypassing the need for manual segmentation in the data synthesis pipeline.

2.5 Classical Computer Vision and Image Processing

While Deep Learning provides powerful data-driven feature extraction, it relies heavily on the availability of ground truth to learn object boundaries. In scenarios where labeled data is scarce or noisy, classical computer vision techniques offer the advantage that they are deterministic and do not require training. In this thesis, established methods, such as graph-based segmentation and edge detection, are utilized as unsupervised structural priors. By leveraging their ability to extract geometric and textural features based on pixel intensity alone, the learning process can be guided, refining noisy pseudo-labels and generating synthetic data without incurring the cost of manual annotation.

2.5.1 Superpixel and Graph-based Segmentation

Prior to the advent of end-to-end semantic segmentation networks, image segmentation was often formulated as a graph partitioning or clustering problem. These methods are particularly valuable for refining noisy pseudo-labels, as they adhere to natural image boundaries, such as edges and color gradients, without requiring training data.

Simple Linear Iterative Clustering (SLIC)

Supapixel algorithms group pixels into perceptually meaningful atomic regions that share similar visual properties. SLIC [2] adapts the k -means clustering algorithm to generate these superpixels. It operates within a five-dimensional space $[l, a, b, x, y]$, integrating color information from the CIELAB color space with spatial coordinates (x, y) . By clustering pixels based on color similarity and spatial proximity, SLIC adheres to image boundaries more effectively than rigid grid structures. In Chapter 3, SLIC superpixels are employed to refine coarse segmentation masks based on their overlap.

GrabCut

GrabCut [126] is an iterative segmentation technique that models the image foreground and background using Gaussian Mixture Models (GMMs). The algorithm represents the image as a graph wherein pixels serve as nodes and edge weights reflect the similarity between neighboring pixels. The segmentation problem is formulated as an energy minimization task and solved using Graph Cuts via the max-flow/min-cut algorithm. GrabCut can be initialized with a bounding box around the target object and, optionally, with a specific mask containing four flag types: Definite Background, Probable Background, Definite Foreground, and Probable Foreground. The algorithm iteratively updates the color models and assigns the "probable" pixels to the most likely category. This capability is utilized in Chapter 3 to refine coarse instance pseudo-masks.

2.5.2 Edge Detection

Edge detection is a fundamental operation used to identify significant intensity discontinuities within an image. These discontinuities typically correspond to physical boundaries or abrupt changes in surface reflectance. While deep neural networks extract such features implicitly, explicit edge detection provides a deterministic framework for structural analysis. In this thesis, edge detection is used as a proxy for texture analysis, where the density and distribution of edges serve as a signature for surface integrity.

Canny Edge Detection

The Canny edge detector [17] is a multi-stage algorithm designed to optimize the trade-off between edge detection and localization. It is particularly valued in computer vision pipelines for its ability to produce precise, single-pixel contours. The algorithm follows a sequential four-stage pipeline:

1. **Gaussian Smoothing:** The input image is convolved with a Gaussian filter to suppress high-frequency noise. This preprocessing step prevents the gradient operator from reacting to minor intensity fluctuations caused by sensor noise.
2. **Gradient Computation:** The algorithm calculates the magnitude and orientation of intensity gradients across the image. This step identifies areas of rapid brightness change, which are primary candidates for edges.

3. Non-Maximum Suppression (NMS): To achieve thinned edges, the algorithm suppresses all gradient responses except for those representing local maxima in the gradient direction. This process converts broad transitions into sharp, single-pixel lines.
4. Hysteresis Thresholding: A dual-threshold approach (T_{high} and T_{low}) is applied to resolve contour ambiguity. Pixels above T_{high} are designated as strong edges, while those between T_{low} and T_{high} are considered weak edges. The latter are preserved only if they are spatially connected to a strong edge. This strategy maintains contour continuity while effectively filtering out spurious noise responses.

In Chapter 5, this classical operator is extended to isolate anomalous surface textures in table grape berries.

Part I

Precision Agriculture

Chapter 3

Bridging Covariate Shift via Pseudo-Labeling

Building upon the environmental and perceptual challenges outlined in Chapter 2, this chapter addresses the core issue of the annotation bottleneck in unstructured agricultural domains. Specifically, it explores how surrogate supervision can be leveraged to mitigate the severe covariate shifts encountered when adapting models from available source datasets to new, unconstrained target fields.

In the field of precision agriculture, the detection and segmentation of fruits and vegetables are essential tasks for applications such as robotic harvesting and yield estimation. As in other automation domains, the performance of these perception systems benefits significantly from controlled environments and predictable field conditions. However, the complexity and reliability of learning-based models are primarily dictated by the degree of environmental uncertainty. Factors such as varying illumination, fruit occlusion, diverse crop varieties, and inconsistent spatial positioning present substantial challenges for standard computer vision algorithms.

To mitigate these challenges, many existing systems are explicitly designed to minimize sources of variability through hardware engineering. For example, [151] and [104] proposed detection systems for fruit counting and yield estimation using direct illumination devices to control ambient light. In both cases, the proposed systems require nighttime operation to remain functional. Another common approach involves the use of straddle robotic platforms, as demonstrated in [118], where the platform physically shields the crop from sunlight, thereby reducing variabilities caused by camera intrinsics and natural lighting.

While these hardware-centric approaches are viable in certain contexts, they are challenging to adapt across different cultivations, such as the overhead trellis systems used for table grapes. The lack of flexibility, coupled with the significant financial and technical investment required, represents a prohibitive barrier for small and medium-sized agricultural businesses. Although these enterprises face the same labor shortages and economic pressures as larger companies, they lack the resources to re-engineer their cultivation methods from the ground up to support hardware-intensive automation. Consequently, there is a growing demand for flexible, software-driven, and data-oriented solutions that can operate in unmodified environments using standard, low-cost sensors. This shift would positively impact these businesses

by making precision agriculture more accessible and enabling a wider range of data-driven applications.

Modern computer vision algorithms, particularly those based on deep learning, offer a promising path to lower this entry barrier. However, deep learning algorithms introduce a new challenge as they are inherently "data-hungry". Gathering the correct type and amount of annotated data is rarely a straightforward process, especially when dealing with the high variability of specific crops such as table grapes [75].

One effective strategy to address data scarcity is the application of semi-supervised, weakly-supervised, and transfer learning techniques, where additional information is incorporated into the training process via external training signals or by leveraging knowledge from a related task. An example of these approaches in agriculture is provided in [13], where an olive counting solution is explicitly trained using weak labels and consistency losses. The auxiliary signal in their case is the labeling obtained by an external classifier that simply detects the presence or absence of olives in the picture. While effective for data with minimal labeling, this method is limited to detection and counting. Such methods do not recover the spatial granularity and shape information provided by instance segmentation, which is required for advanced robotic tasks such as manipulation or for precise volume estimation. Similarly, transfer learning can be used to reduce the amount of data needed for training. [50] demonstrated that unsupervised contrastive pre-training significantly improves detection performance compared to standard ImageNet pre-trained initialization. Although transfer learning provides a strong starting point for training a deep neural network, it does not inherently solve the alignment problem when the target domain differs significantly from the source data, and task-specific data is still required.

In this context, it is important to consider the specific challenges perception algorithms face when deployed in fields that are not explicitly prepared for automation. Factors such as uneven fruit distribution, intra-species variability, shifting illumination, heavy occlusion, and background clutter introduce substantial noise into the system. From a technical perspective, these factors translate into severe covariate shifts and a persistent lack of representative labeled samples. For instance, [51] conducted a study on the intra-species covariate shift in sweet peppers and its impact on detection and segmentation algorithms. The authors explored generalizability by analyzing sweet peppers grown using various cultivars and in different environments (open fields versus glasshouses). Their results show that single-task learning performance drops significantly during cross-dataset detection. Only by setting up a multi-task learning problem were they able to increase their performance margins, leveraging the multiple back-propagation signals. Exploiting cross-task correlations can be viewed as a form of self-supervised learning. Nevertheless, models trained on a specific source distribution often fail to generalize to a target environment without explicit adaptation, and it is often difficult to collect a sufficient number of labeled images that accurately capture the actual field variability.

The grapevine is a prime example of a crop with a wide range of varietal variability. Wine and table grapes differ significantly in size and bunch structure, and the vines are trained using completely different structural systems. Even among different varieties of the same grape type, the variations in size, shape, color, foliage,

and vine structure make it difficult for detection and segmentation algorithms to generalize. Early approaches to grape detection and counting primarily relied on the use of fine-tuned hand-crafted features. For example, [132] used Histogram of Oriented Gradients (HOG) descriptors together with a Support Vector Machine (SVM) to build a white wine grape detector. Building on those early results, [119] proposed a solution based on HOG, Fast Radial Symmetry Transform (FRST), and Linear Binary Patterns (LBP) to feed an SVM-based detector, using geometrical considerations to separate self-occluding grape bunches. Yield estimation in such frameworks is typically based on counting the number of detected berries. While both solutions show some robustness to color and illumination variability, they require substantial manual tuning of the algorithms, which limits the transferability of trained systems to other grape varieties. Another approach to grape yield estimation based on geometrical considerations is presented in [92], where detection is performed in an unsupervised manner on early-stage buds sprouting from the branches using Gaussian fitting. The benefits of this approach lie in its independence from labeled data and its reliance on a simple camera. However, this method is only applicable in the early season and does not account for yield loss due to malformed or diseased grape bunches.

More recently, numerous deep learning-based detection algorithms have been proposed. [108] presented an early yield prediction system based on berry counting using a SegNet segmentation network to extract features from bunches and canopy images, such as the estimated number of visible berries and the ratio of leaves within detection bounding boxes. Their statistical analysis of feature variability across six grapevine varieties highlights how cross-variety distribution shifts represent a significant limitation for machine learning techniques. Furthermore, their system operates exclusively at night under direct illumination. Another berry detection system based on CNNs is presented in [165] and in the extended work [166]. The authors achieve excellent accuracy results for two grapevine training systems, but their images were collected using a modified straddle harvester to maintain consistent illumination. This hardware requirement makes the solution inapplicable to table grapes, where straddle machines cannot be used due to the overhead trellis structure. In [29], a counting network inspired by density-based crowd counting techniques was presented, achieving impressive mean average errors. Attaining these results, however, required manual labeling of over 35,000 berries, which is an extremely time-consuming process. All these approaches leverage deep learning for increased performance but limit the variability of the problem by relying either on highly specialized machinery or on massive quantities of manually labeled data. These aspects limit the generalizability and applicability of such techniques beyond their original use cases.

In contrast to these data-intensive or hardware-restricted approaches, this chapter proposes a methodology for training high-performance detection and instance segmentation algorithms with minimal labeled data. This chapter explicitly focuses on a covariate shift scenario in which only a small amount of labeled data from a similar cultivation (source dataset) is available, which is insufficient to achieve acceptable performance in a different orchard (target dataset). In particular, the source dataset is the Embrapa Wine Grape Instance Segmentation Dataset (WGISD) presented in [127], while the target data consists of Pizzutello table grapes cultivated

in Aprilia (southern Lazio, Italy), which exhibit significant differences compared to the source data.

To bridge this gap without expensive manual annotation, a unified system is presented based on the core concept of surrogate supervision. In the context of this chapter, surrogate supervision refers to the strategy of deriving structured training signals from weak cues and structural priors rather than relying on dense manual annotations. This is achieved by combining weakly-supervised and semi-supervised techniques to generate pseudo-labeled data for both detection and instance segmentation, addressing the covariate shifts that arise when a new crop becomes the target of a computer vision system for precision agriculture. The approach is designed to operate with limited hardware, using simple images captured with a standard smartphone camera and videos collected with an Intel RealSense D435i. Although the latter is a depth-capable sensor, only the RGB stream was utilized in this study, demonstrating the methodology’s feasibility using accessible consumer-grade hardware for small and medium-sized agricultural businesses.

Specifically, the proposed pseudo-labeling strategies focus on two main contributions:

1. **Automatic bounding box generation:** A method for propagating bounding box annotations across consecutive video frames based on a sparse initial labeling and geometric feature matching. The system automatically generates high-quality pseudo-labels for unannotated frames, significantly increasing the performance of the target detector.
2. **Pseudo-mask generation for instance segmentation:** An approach that employs bounding box annotations, either manually annotated or automatically generated by the aforementioned detection system, to spatially guide the predictions of an instance segmentation model initially trained only on the source dataset. However, simply re-training on these initial predictions would lead to confirmation bias, reinforcing the model’s existing errors. To overcome this, several mask refinement strategies are implemented and compared, employing morphological dilation, SLIC superpixels, and GrabCut. The resulting high-quality masks are then used as a training signal for the final instance segmentation model, which is trained on both the source and target datasets.

The system is validated in two distinct scenarios. First, ground truth bounding boxes are used to isolate the segmentation performance, establishing the theoretical upper bound of the system. Second, bounding boxes predicted by the external detector are used to demonstrate the system’s capability to operate as a fully automated, end-to-end pipeline for precision agriculture.

3.1 Materials and Methods

This Section details the data collection process, the overall system framework, and the specific algorithms employed to generate pseudo-labels for detection and instance segmentation tasks.

3.1.1 Target and Source Datasets

The experimental activities were conducted in the commercial vineyard located in Aprilia (Italy), the details of which are described in Section 2.1.3.

The vineyard contains four table grape varieties: White Pizzutello, Black Pizzutello, Red Globe, and Black Magic (Figure 3.1). This study focuses on the Black Pizzutello variety (Figure 3.1d), which is specific to the Lazio region and presents significant intra-class variability. Unlike standard spherical berries, Pizzutello berries possess elongated and irregular shapes. Furthermore, the Black Pizzutello shares identical shape characteristics with the White variety and, during early ripening stages, exhibits a similar color. This high variability in shape and color makes it an ideal candidate for testing the robustness of detection and instance segmentation algorithms against covariate shift. The proposed system addresses the covariate shift from a generic source dataset to a target dataset representative of actual field conditions.

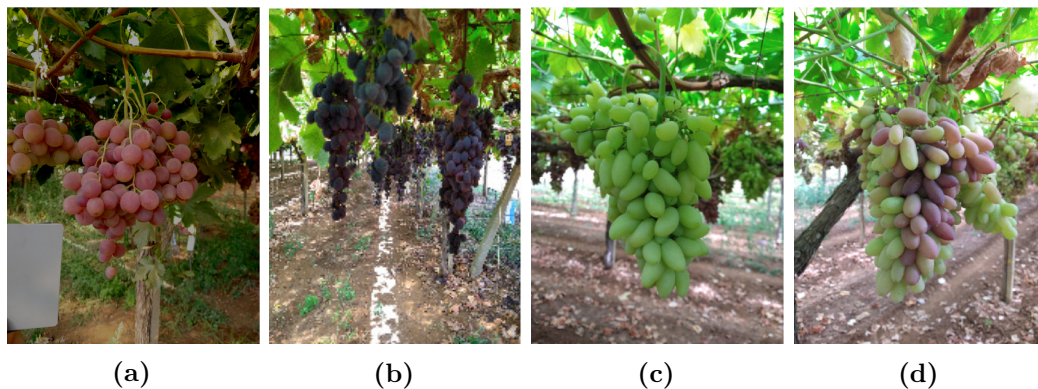


Figure 3.1. Representative examples of the four table grape varieties cultivated in the experimental vineyard. The Black Pizzutello variety (d) serves as the primary focus of this study due to its pronounced intra-class variability in both shape and color compared to standard spherical berry variants.

The target dataset is composed of two distinct data types: a video dataset and a static image dataset. The Target Video (TVid) dataset comprises high-definition (1280×720) videos recorded at 10 Hz using an Intel RealSense D435i camera. The camera was moved along the vineyard rows without strict requirements regarding distance from the fruits or height from the ground, simulating a data collection operation that could be easily performed by a farmer or a mobile robotic platform. A total of 1469 frames were extracted from these video sequences, with examples illustrated in Figures 3.2a and 3.2b. Although a neutral gray card was utilized during data collection for white balance purposes, this was intended for potential future research and does not constitute a requirement for the methodology presented in this chapter. A short 10-second segment was manually labeled for testing purposes, while the remainder of the dataset was utilized without ground truth labels, leveraging the semi-supervised nature of the proposed system.

The Target Image (TImg) dataset consists of 134 static images at 3000×4000 resolution collected using a standard smartphone camera (Moto G8 Plus). This data simulates the images a farmer or a robot could collect to perform agricultural

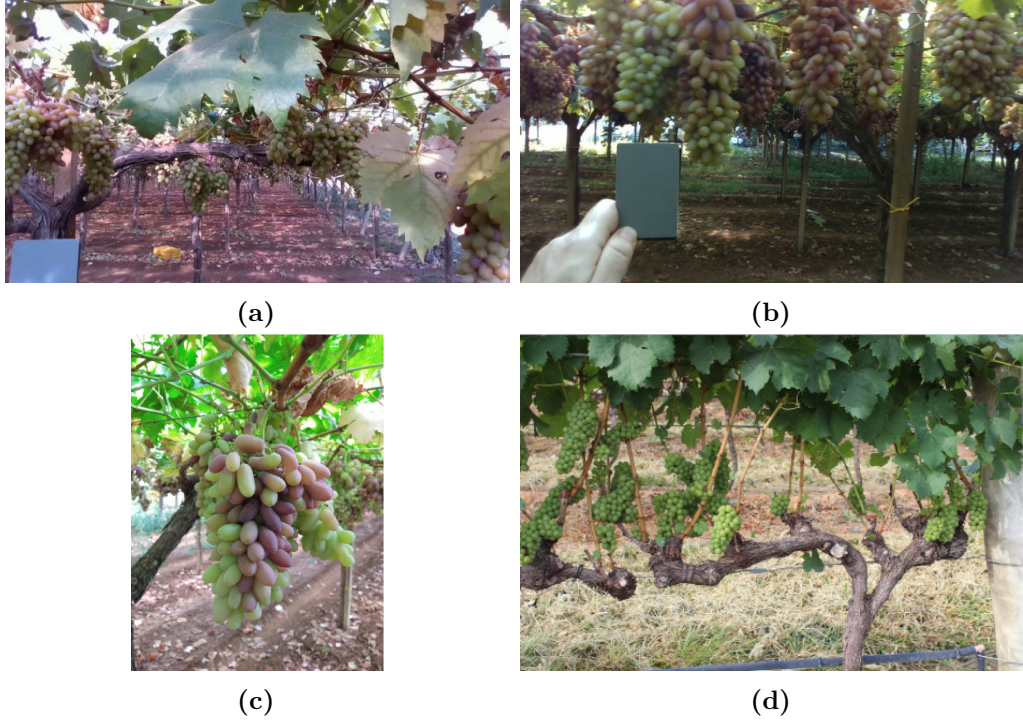


Figure 3.2. Visual comparison illustrating the covariate shift between the source and target datasets. (a, b) Representative frames from the target video dataset (TVid), captured using a standard consumer-grade camera to simulate unconstrained field acquisitions. (c) A static image from the target image dataset (TImg). (d) An example from the source dataset (WGISD) [127]. The significant variations in berry morphology, color distribution, and illumination conditions between the source and target domains highlight the necessity for domain adaptation techniques.

actions on specific grape bunches, such as quality estimation, disease detection, or automatic harvesting. Utilizing different optics, sensors, and capturing modes for the video and still images introduces a common source of covariate shift, further testing the robustness of the system. All images in this set were manually labeled with bounding boxes for object detection. Additionally, a subset of 70 images was also annotated with precise instance segmentation masks and were used as the validation and test sets.

The source dataset serves as the initial training ground for the models. In this study, the Embrapa Wine Grape Instance Segmentation Dataset (WGISD) [127] is utilized. The WGISD introduces a substantial covariate shift compared to the target data in terms of grape variety (wine versus table grapes, berry shape, and color), training structure, illumination conditions, camera devices, and image scales.

3.1.2 System Overview

The proposed methodology addresses the annotation bottleneck challenge by leveraging surrogate supervision to bridge the covariate shift between different agricultural domains. As illustrated in Figure 3.3, the system framework operates in two interconnected subsystems: the Detection Pseudo-Label Generation (DPLG) and

Segmentation Pseudo-Label Generation (SPLG).

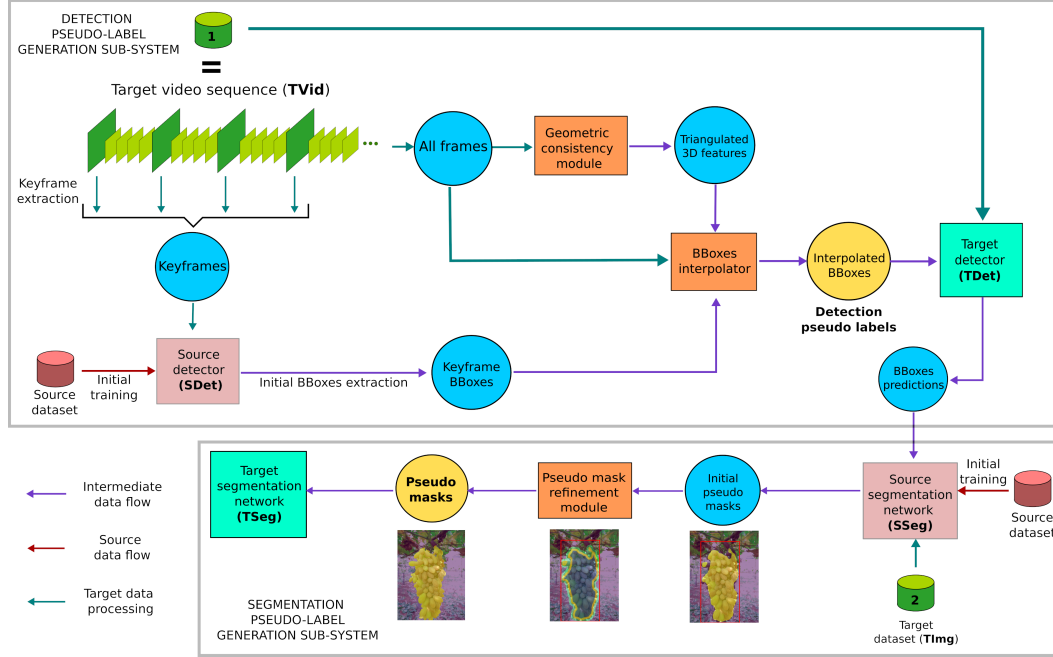


Figure 3.3. Overall architecture of the proposed surrogate supervision framework. The pipeline operates in two primary stages to bridge the domain gap. First, the Detection Pseudo-Label Generation (DPLG) subsystem utilizes a source-trained detector (SDet) on sparse video keyframes to initialize geometric feature propagation, training the target detector (TDet) with interpolated bounding boxes derived from the target video sequence (TVid). Subsequently, the Segmentation Pseudo-Label Generation (SPLG) subsystem leverages TDet to generate bounding boxes on a target dataset (e.g., TImg), which are used as spatial priors for a source-trained segmentation network (SSeg). The resulting initial masks undergo a refinement process to generate high-quality pseudo-masks for training the final target segmentation model (TSeg).

Within the DPLG subsystem, the source dataset is used to train the initial detection model, referred to as the Source Detector (SDet) network. Because SDet is not optimized for the target environment, its direct application yields suboptimal results. However, it can be applied to a selected subset of frames from the TVid dataset, designated as keyframes. Initial bounding boxes are extracted from these keyframes using a high confidence threshold to limit false positives. Subsequently, all frames of the video sequence are processed by a Geometric Consistency module that extracts and associates visual features across consecutive frames. This geometric information, combined with the initial bounding boxes from the keyframes, enables the highly accurate propagation of bounding box positions to the remaining unannotated frames. The combination of the initial bounding boxes and the propagated ones serve as detection pseudo-labels to train the Target Detector (TDet).

Following this, the SPLG subsystem aims to produce accurate pixel-level masks. The source dataset is first used to train the Source Segmentation (SSeg) network. While SSeg struggles to generate precise masks on the target data independently, its performance can be significantly improved by providing spatial cues. The bounding

box predictions generated by the TDet (or manually annotated bounding boxes) on the TImg dataset are utilized as spatial priors to guide SSeg, which outputs initial pseudo-masks. To prevent confirmation bias during subsequent training, these initial pseudo-masks are processed through a refinement module utilizing classical computer vision techniques. The resulting refined pseudo-masks are finally used to train the Target Segmentation (TSeg) network on the target distribution.

3.1.3 Detection Pseudo-Label Generation

The DPLG subsystem, depicted in Figure 3.4, relies on a Geometric Consistency module to associate grape instances across different frames of a video stream. By identifying 2D features belonging to a detected grape bunch and matching them across consecutive frames, the same instance can be accurately mapped over time.

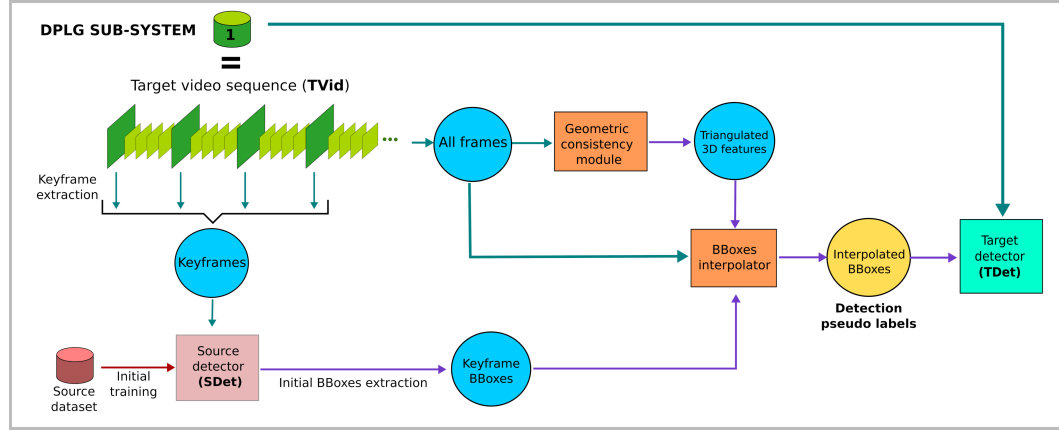


Figure 3.4. Workflow of the Detection Pseudo-Label Generation (DPLG) subsystem. Bounding box predictions generated by the source detector (SDet) on video keyframes are propagated to intermediate unannotated frames using a geometric consistency module, providing a dense set of surrogate bounding box labels for training the target detector (TDet).

Because the videos collected in the vineyard consist of simple linear trajectories without closed loops, the visual displacement between consecutive frames captured at 10 Hz is minimal, and each table grape bunch appears only in a limited sequence of consecutive frames. Therefore, extracting 2D features from a given frame i and matching them with a small number of subsequent frames $i + 1, \dots, i + n$ is sufficient to map the grape instance along the video stream. Speeded Up Robust Features (SURF) [11] and their corresponding descriptors are extracted, and matching is performed using a brute-force distance computation between descriptors.

To filter out mismatched features, a Random Sample Consensus (RANSAC) [42] verification step is applied. Assuming that the image transformation between consecutive frames with small camera motion can be approximated by a homography, four matches are repeatedly sampled to compute a relative homography matrix. The homography with the highest consensus is selected, and all matches whose spatial displacement is incompatible with this transformation are discarded as outliers. An example of the feature matching before and after this verification is given in Figure

3.5.

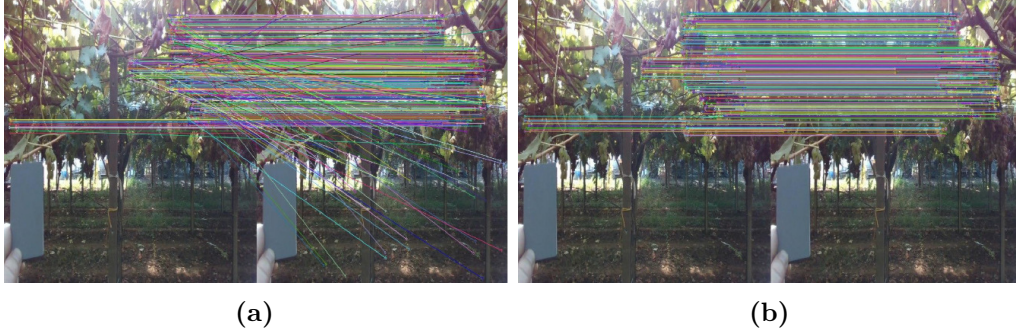


Figure 3.5. Geometric verification process for robust instance association across consecutive video frames. (a) Initial 2D feature correspondences established via brute-force matching of SURF descriptors. (b) Refined feature matches following homography estimation and RANSAC-based outlier rejection, ensuring the reliable spatial propagation of bounding boxes despite camera motion.

The bounding box propagation process leverages this filtered feature matching. The system employs an alternating prediction and propagation strategy. Specifically, SDet predicts the bounding boxes of every other frame (e.g., frame i and frame $i + 2$). For the intermediate unannotated frame (frame $i + 1$), the Geometric Consistency module identifies the associated 2D features inside the bounding box predicted by SDet at frame i and matches them to features in frame $i + 1$. Under the assumption that the camera is moving slowly and tangentially to the vineyard row, the physical scale of the bunch changes minimally over a single frame interval. Therefore, the new bounding box in frame $i + 1$ is assigned the same dimensions as the box in frame i . The spatial position of the new bounding box is updated by calculating the center of gravity of the matched features in frame $i + 1$ and aligning the center of the bounding box to this point. This process is shown in Figure 3.6.

The combination of the bounding boxes predicted by SDet and those generated via geometric propagation acts as the final set of detection pseudo-labels used to train the Target Detector (TDet), completing the DPLG pipeline. For the detection subsystem, the single-stage YOLOv5 architecture was utilized. Single-stage detectors extract region proposals and class predictions in a single pass, offering high-speed inference suitable for future real-time applications on constrained hardware.

3.1.4 Segmentation Pseudo-Label Generation

The SPLG subsystem, depicted in Figure 3.7, encapsulates the process of adapting the SSeg network using bounding box cues on the target data, refining the initial predictions, and training the final TSeg network.

For segmentation and pseudo-mask generation, the SSeg model utilizes a two-stage Mask R-CNN architecture equipped with a ResNet-101 [56] backbone. This model is initially pre-trained on the WGISD source dataset. To effectively generate pseudo-labels without requiring dense manual annotations on the target dataset, we deviate from its standard inference procedure. The specific modifications and refinement steps are detailed below.

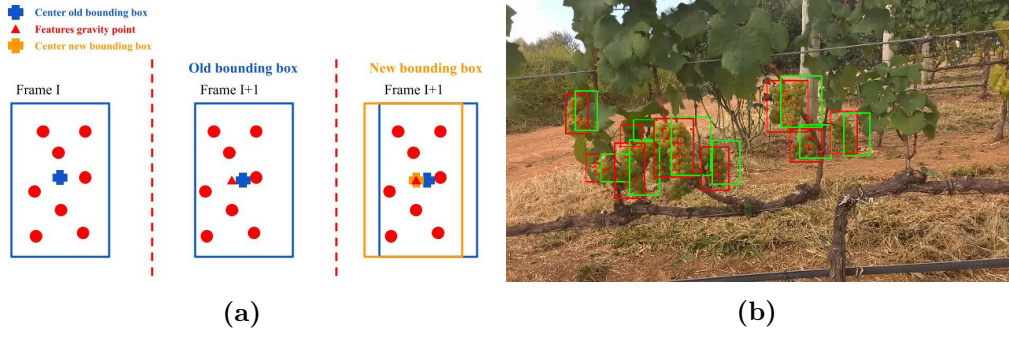


Figure 3.6. The bounding box propagation mechanism. (a) Schematic of the updating principle: the bounding box predicted in frame i (blue) is projected onto frame $i + n$. Assuming negligible scale variation, the dimensions are preserved while the spatial coordinates are updated by aligning the box center with the newly computed center of gravity (orange) of the features. (b) Visualization of the generated pseudo-labels. Green boxes denote the original predictions from the source detector (SDet) at frame i , while red boxes represent the geometrically propagated labels in frame $i + 1$ based on the matched features (red points).

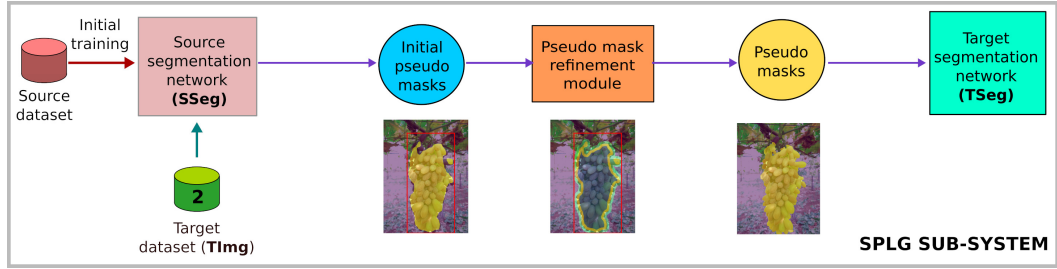


Figure 3.7. Workflow of the Segmentation Pseudo-Label Generation (SPLG) subsystem. The source-trained segmentation network (SSeg) processes the target dataset (TImg) utilizing external spatial priors to generate initial pseudo-masks. These coarse masks are subsequently processed through a refinement module to produce high-quality pseudo-labels for training the target segmentation network (TSeg).

Standard Mask R-CNN Workflow

In its standard formulation, Mask R-CNN operates in two stages. First, a Region Proposal Network (RPN) generates candidate Regions of Interest (RoIs) likely to contain objects. Second, these RoIs are processed by two parallel heads: a bounding box head for classification and coordinate regression, and a mask head for pixel-wise segmentation.

However, the data flow differs between training and inference. During training, the bounding box head and the mask head operate in parallel on the proposals generated by the RPN. This parallel design allows the mask head to learn robust segmentation even from imperfect, unrefined proposals. Conversely, during standard inference, the process is sequential. The RPN generates proposals, which are first refined by the bounding box head to produce tight, filtered, high-confidence detections. The localized image features corresponding to these refined bounding boxes are then extracted and passed to the mask head to generate the final segmentation.

This distinction is depicted in Figure 3.8.

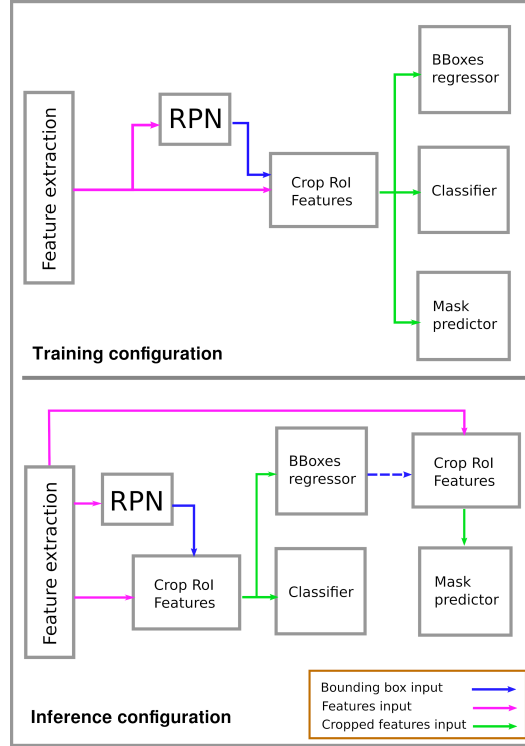


Figure 3.8. Data flow within the Mask R-CNN architecture. (Top) During training, the bounding box and mask heads operate in parallel on proposals generated by the Region Proposal Network (RPN). (Bottom) During standard inference, the process is sequential, with the mask head processing localized features based on the refined bounding boxes predicted by the detection branch. To generate pseudo-labels without confirmation bias, the region proposal and bounding box regression stages are bypassed. The dashed blue line illustrates the injection point where external bounding box predictions are provided directly as spatial priors for the mask head.

Modified Workflow for Pseudo-Labeling

To mitigate the confirmation bias that would occur if SSeg relied on its own potentially flawed detections in the target domain, our system implements a modified inference pipeline that bypasses the detection stage entirely. We drop the bounding box regression and instead inject external bounding boxes directly into the mask branch.

In this configuration, the external bounding boxes (either manually annotated ground truth or generated by the DPLG subsystem) serve as strong spatial priors. They are used to extract localized features, which are then passed to the mask head. This effectively forces the network to segment the area defined by the input box, regardless of its own classification confidence. This modified flow offers two key advantages. First, by deriving bounding boxes from an external source, we effectively break the confirmation bias loop. Second, we can isolate the performance contribution of the mask generation process by injecting ground truth bounding boxes. In this setting, the number of pseudo-masks exactly matches the actual

number of grape clusters, and any error in the generated pseudo-labels can be attributed solely to the mask generation process, isolating it from detection errors. The qualitative difference in the segmentation mask between the standard Mask R-CNN wiring and that with an external spatial prior is shown in Figure 3.9.



Figure 3.9. Impact of external spatial priors on segmentation quality in the target domain. (Left) Standard inference output from the model trained exclusively on source data, demonstrating a failure to accurately localize and segment the target instance. (Right) Output from the identical model utilizing the guided inference strategy, demonstrating how an external bounding box cue forces the network to appropriately attend to the relevant spatial region.

Pseudo Mask Refining Block

To further reduce confirmation bias and improve mask quality on the target data, the generated pseudo-masks are refined using classical computer vision techniques. Three distinct strategies relying on geometrical and color considerations were implemented and compared:

- **Morphological Dilation:** The SSeg model tends to underestimate mask boundaries on the target data. To address this, a simple morphological dilation expands the initial pseudo-mask until it touches the boundaries of the reference bounding box, adding valuable spatial information to the label. The dilation is applied using a 5×5 circular-shaped kernel (Figure 3.10a).
- **Simple Linear Iterative Clustering (SLIC):** SLIC [2] is a super-pixel segmentation method that clusters contiguous regions of an image using a k-means approach based on color and spatial coordinates (a 5-dimensional space). Using the Python scikit-image library [144], the image is divided into 2000 segments with compactness set to 0.1. These super-pixels are then compared

against each initial pseudo-mask. Super-pixels covered by the mask above an upper threshold ($t_u = 70\%$) are added to the final mask, while those covered below a lower threshold ($t_l = 30\%$) are removed. This strategy effectively refines the initial prediction, aligning the mask with natural image boundaries and removing erroneously included background pixels (Figure 3.10b).

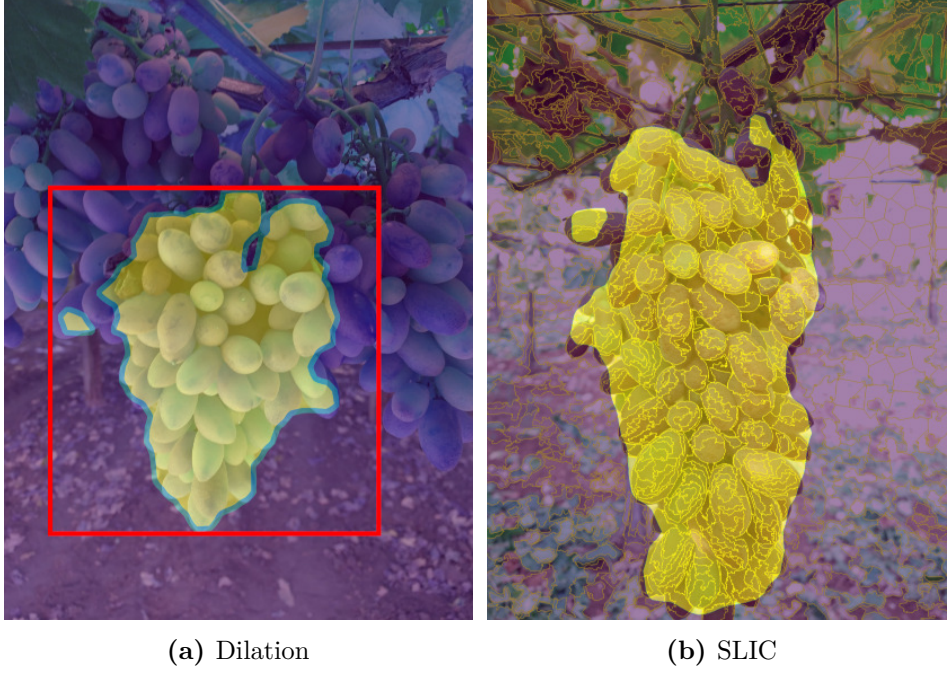


Figure 3.10. Examples of mask refinement strategies applied to initial pseudo-masks. (a) Morphological dilation, where the yellow region indicates the initial pseudo-mask and the green boundary illustrates the spatial expansion achieved by the dilation operation to capture missing boundary pixels. (b) Super-pixel segmentation utilizing the Simple Linear Iterative Clustering (SLIC) algorithm, which groups pixels based on color and spatial proximity to align the mask with natural image boundaries.

- **GrabCut:** This iterative segmentation technique [126] represents the image as a graph where foreground and background pixels are modeled as Gaussian Mixture Models separated iteratively by graph cuts. Using the OpenCV [16] implementation, the algorithm is initialized with a mask defining four pixel categories: sure foreground, sure background, probable foreground, and probable background. The initial pseudo-mask is designated as the probable foreground. Morphological dilation is applied to the pseudo-mask for a number of iterations proportional to the smallest dimension of the reference bounding box to obtain the probable background. Erosion is applied for the same number of iterations to obtain the sure foreground, while the remaining pixels are set to sure background (Figure 3.11).

The resulting refined pseudo-masks are subsequently utilized as ground-truth labels to train the final TSeg network, concluding the SPLG pipeline.

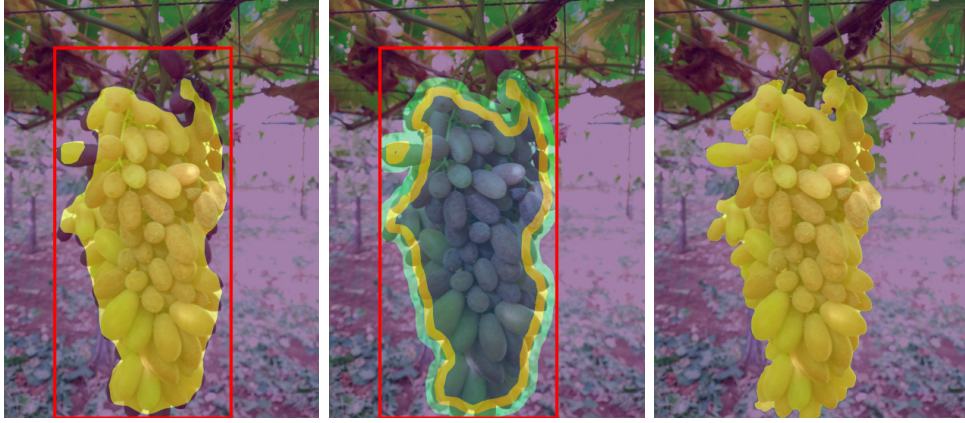


Figure 3.11. Visual example of a pseudo-mask refined using the GrabCut algorithm. The colored overlays denote the initial pixel assignments: sure foreground (blue), probable foreground (yellow), probable background (green), and sure background (purple).

3.1.5 Metrics

The performance of the detection and segmentation algorithms is evaluated using standard computer vision metrics: Precision, Recall, Intersection over Union (IoU), and Average Precision (AP). Because the dataset contains only a single class (grapes), the Average Precision (AP) coincides with the mean Average Precision (mAP). Following the MS COCO [89] evaluation standards, AP is calculated by computing the precision at every recall level from 0 to 1 with a step size of 0.01. The primary metric, $mAP_{0.5:0.95}$, is computed by averaging the AP over ten IoU thresholds ranging from 0.50 to 0.95 with a step size of 0.05. Additionally, $mAP_{0.50}$, $mAP_{0.75}$, and $mAP_{0.95}$ are reported to evaluate performance at specific overlap thresholds.

3.2 Results and Discussion

This section presents the experimental validation of the proposed framework. The evaluation is divided into three parts. First, the detection subsystem is assessed to quantify the improvements gained through geometric pseudo-labeling. Second, the segmentation subsystem is evaluated in isolation to analyze the impact of different mask refinement strategies. Finally, the complete end-to-end pipeline is tested, demonstrating the system’s ability to autonomously generate high-quality segmentation models from unlabeled video data.

3.2.1 Object Detection Experiments

The initial phase of the evaluation established the performance baseline for the target domain. To reduce the risk of overfitting and to ensure the resulting system remains viable for deployment on edge devices typical in robotic harvesting, the lightweight YOLOv5s (Small) architecture was selected as the core detection engine for all experiments.

All detection models were trained using an NVIDIA DGX workstation. The training protocol consisted of 300 epochs with a batch size of 4 and an early stopping

patience of 30 epochs. A one-cycle learning rate scheduling strategy was employed, with an initial learning rate of 0.01 and a final one of 0.001. The models were optimized using Stochastic Gradient Descent (SGD) with a momentum of 0.937 and a weight decay of 5×10^{-4} . The networks were initialized with MS COCO [89] pretrained weights and subsequently fine-tuned. To encourage generalization, the 242 training images of the source dataset were augmented offline using random cropping, random contrast adjustments, Gaussian blur, Gaussian noise, and horizontal flipping. This process generated 726 additional augmented images, resulting in a total source training set of 968 images.

To assess the effectiveness of the DPLG subsystem, the baseline SDet, which was trained exclusively on the source data, was compared against the TDet, which was fine-tuned on a combination of the source dataset and the automatically generated pseudo-labels from the target domain.

Table 3.1 shows the performance on the static target images (TImg). Incorporating the pseudo-labels generated from the videos increased the $mAP_{0.5}$ from 0.69 to 0.77. This improvement is particularly notable considering that the static images and the video frames possess different underlying distributions due to varying optics, camera sensors, and capture modes.

Table 3.1. Performance comparison between the source detector (SDet) and the target detector (TDet) on the test split of the static image target dataset (TImg).

Model	Precision	Recall	$mAP_{0.5}$	$mAP_{0.95}$
SDet	0.90	0.56	0.69	0.46
TDet	0.98	0.68	0.77	0.47

A qualitative visual comparison is provided in Figure 3.12. The upper row illustrates the predictions made by SDet, while the lower row shows the output from TDet on the same images. The TDet model not only produces tighter bounding boxes around the grape clusters but also successfully detects a higher number of instances, reflecting the quantitative improvements in both precision and recall.

Because the TImg and TVid datasets represent different operational modalities, TDet was additionally evaluated on the test sequences of the TVid dataset. The results, summarized in Table 3.2, reveal a gain of 0.10 in $mAP_{0.5}$ for the network trained with pseudo-labels compared to the baseline. The more pronounced performance gain in this scenario is expected, as the covariate shift between the TVid training and test data is minimal compared to the shift between the video and static image domains.

Table 3.2. Performance comparison between the source detector (SDet) and the target detector (TDet) on the test split of the video target dataset (TVid).

Model	Precision	Recall	$mAP_{0.5}$	$mAP_{0.95}$
SDet	0.62	0.59	0.55	0.21
TDet	0.74	0.60	0.65	0.23



Figure 3.12. Qualitative comparison of object detection performance before and after domain adaptation. The top row illustrates the predictions generated by the baseline source detector (SDet), which struggles with missed instances and loosely fitted bounding boxes. The bottom row displays the predictions from the target detector (TDet) trained with surrogate supervision, demonstrating tighter localizations and significantly higher recall, particularly in dense and occluded canopy clusters.

3.2.2 Instance Segmentation Experiments

The segmentation models were built upon the Mask R-CNN architecture with a ResNet-101 [56] backbone and implemented using the Detectron2 framework [157]. Training was conducted on the NVIDIA DGX cluster starting from MS COCO pretrained weights. Standard data augmentation techniques were applied, including Gaussian blur, Gaussian noise, random brightness and contrast variations, pixel dropout, random flipping, and random cropping. The networks were optimized using SGD with a learning rate of 0.001, a weight decay of 0.0001, and a momentum of 0.9 for a maximum of 100 epochs, utilizing early stopping monitored against the validation AP with a patience of 20 epochs.

To establish the severity of the covariate shift, the baseline SSeg network was evaluated directly on the TImg dataset. To mitigate the obvious scale differences between the source and target domains, cropping and resizing augmentations were applied to the source dataset during training. The results in Table 3.3 show a dramatic performance drop of over 20 points in Average Precision across all metrics when transitioning from the WGSD test set to the TImg test set. This drop highlights the necessity of domain adaptation techniques.

To evaluate the Segmentation Pseudo-Label Generation (SPLG) subsystem independently of detection errors, an ablation study was conducted using ground truth bounding boxes. This isolated setup ensures the number of generated pseudo-

Table 3.3. Quantitative evaluation of the covariate shift impact on the baseline Mask R-CNN model (SSeg), trained exclusively on the source dataset (WGISD). The model suffers a severe performance degradation when evaluated on the target dataset (TImg)

Test data	Task	$mAP_{0.5:0.95}$	$mAP_{0.5}$	$mAP_{0.75}$
WGISD	Detection	53.40	87.02	57.36
	Segmentation	53.60	89.44	55.41
TImg	Detection	32.65	60.40	30.37
	Segmentation	32.88	65.40	34.77

masks matches the exact number of grape clusters, allowing the evaluation to focus strictly on the mask generation and refinement quality. The Target Segmentation (TSeg) network was then trained on a combination of the source dataset and the newly pseudo-labeled TImg training set.

Table 3.4 reports the average performance over five trials for each refinement strategy evaluated on the TImg test set. Introducing unrefined pseudo-masks immediately yields a substantial improvement of nearly 50% over the baseline SSeg. Among the refinement strategies, GrabCut proved to be the most effective. The addition of the GrabCut refinement block increased the $mAP_{0.5:0.95}$ by 1.13 points and the $mAP_{0.75}$ by 4.58 points compared to the unrefined pseudo-masks.

Table 3.4. Comparison of pseudo-mask refinement strategies. The target segmentation model (TSeg) was trained using different processing techniques applied to the initial pseudo-masks and evaluated on the TImg test set. The integration of the GrabCut algorithm provides the most substantial improvement.

Training Data	$mAP_{0.5:0.95}$	$mAP_{0.5}$	$mAP_{0.75}$
WGISD (baseline)	32.88	65.40	34.77
WGISD + TImg	48.43	83.06	53.12
WGISD + TImg w/ Dilation	48.67	81.54	54.87
WGISD + TImg w/ SLIC	47.78	80.41	53.54
WGISD + TImg w/ Grabcut	49.56	81.03	57.70

3.2.3 Complete System Experiments

The final phase of the evaluation combined the detection and segmentation subsystems into a fully automated, end-to-end pipeline. In this configuration, the TDet was used to predict bounding boxes for the target images. These predictions served as the external spatial priors for the SSeg network to generate initial pseudo-masks, which were subsequently refined to train the final TSeg models.

This end-to-end generation process was executed on both the training sets of TImg and TVid. The resulting models were then evaluated against the TImg test set. This setup presents a highly realistic scenario where models must generalize across differing target distributions.

As shown in Table 3.5, the fully automated pipeline delivers robust performance gains. Using refined pseudo-labels from the TImg dataset alone improves the baseline $mAP_{0.5:0.95}$ by over 41%. Leveraging the larger volume of video data (TVid) yields

even higher performance, despite the inherent domain shift between the video frames used for training and the static images used for testing. The system achieves an approximate 42% increase in the $mAP_{0.5:0.95}$ over the baseline using only video-derived pseudo-labels.

Table 3.5. End-to-end system evaluation on the TImg test set. The performance of the final target segmentation model (TSeg) is evaluated when trained on various combinations of source and pseudo-labeled target data. The bounding boxes used as spatial priors for the target data were generated by the target detector model (TDet), bypassing manual annotation. The total number of images in each training set is indicated in parentheses.

Training Data	$mAP_{0.5:0.95}$	$mAP_{0.5}$	$mAP_{0.75}$
WGISD (baseline) (88)	32.88	65.40	34.77
WGISD + TImg (182)	44.45	75.68	49.78
WGISD + TImg (182) w/ Grabcut	46.41	78.09	51.74
WGISD + TVid (687)	45.99	78.30	52.59
WGISD + TVid (687) w/ Grabcut	46.66	77.38	52.22
WGISD + TImg + TVid (781)	47.44	76.63	56.06
WGISD + TImg + TVid (781) w/ Grabcut	47.81	77.23	53.27

The maximum performance is attained when combining both TImg and TVid training data. Consistent with the isolated segmentation experiments, the GrabCut refinement block provides a definitive advantage over unrefined labels. The best-performing configuration (combining TImg and TVid with GrabCut refinement) improves the $mAP_{0.5:0.95}$ from the baseline of 32.88 to 47.81. This significant improvement confirms that the proposed end-to-end framework not only detects a higher volume of target instances but systematically improves the pixel-level precision of the resulting masks, effectively mitigating the annotation bottleneck without requiring dense manual supervision.

3.3 Concluding Remarks

This chapter addressed the annotation bottleneck and the challenge of covariate shift in unstructured agricultural environments by proposing a weakly and semi-supervised learning framework. By leveraging surrogate supervision, the proposed solution demonstrated that it is possible to adapt deep learning models from a generic source domain to a specific target domain without the prohibitive cost of dense manual annotation. The pipeline operates in two interconnected stages. First, a detection pseudo-label generation subsystem utilizes a generic source detector to extract initial bounding boxes on video keyframes. These predictions are propagated to unannotated frames via geometric feature correspondence and homography verification, creating a comprehensive set of pseudo-labels to train a target-domain object detector. Second, these robust bounding box predictions are injected as spatial priors into a source-trained segmentation network. This guided inference bypasses the network’s internal region proposals to generate initial pseudo-masks. Finally, these coarse masks are refined using classical computer vision techniques, with GrabCut proving particularly effective at recovering precise instance boundaries for training the final target segmentation model.

The experimental results confirmed that this approach significantly improves both object detection and instance segmentation performance when adapting to specific field conditions and sensor modalities. By relying on simple, consumer-grade video acquisitions rather than expensive expert pixel-level labeling, the framework proves that high-quality perception is achievable even under strictly constrained annotation budgets. This effectively lowers the barrier to entry for deploying data-driven robotic harvesting and monitoring systems in diverse, real-world orchards.

While this pseudo-labeling strategy maximizes the utility of available field data, its generalization capacity remains ultimately bounded by the physical data that can be collected. Relying solely on manual acquisition to cover various vineyard configurations or specific lighting conditions presents a logistical challenge. Consequently, there is a clear need to scale model robustness by shifting the burden of generating diverse environmental contexts to simulation, thereby relaxing the physical constraints of manual data collection. The following chapter addresses this by exploring how simulated environments can be leveraged to automatically generate hybrid training data.

Chapter 4

Automated Hybrid Data Generation for Detection

Developing robust perception models for unstructured agricultural environments requires substantial volumes of training data capable of capturing diverse combinations of lighting, viewpoints, and canopy structures. While various techniques can mitigate the annotation bottleneck, the generalization capacity of these models remains constrained by the physical data that can realistically be acquired. Relying exclusively on manual in-field collection to scale the training distribution presents a significant logistical limitation. To overcome these physical constraints and complement real-world data with synthetic variations, this chapter introduces a compositional synthesis approach.

This shift toward scalable, synthetic data generation is particularly relevant given the rapid advancement of robotics in precision agriculture. The deployment of autonomous field systems is largely driven by the adoption of data-driven algorithms for various monitoring and operational tasks [22, 34]. Within this context, computer vision has become integral for many applications such as yield estimation [77], pest and disease control [78], and visual servoing for harvesting or spraying [118, 90, 136]. However, deploying these systems in the agricultural domain presents unique perception challenges due to the constantly changing environment, which is driven by normal agricultural operations, and the inherent morphological diversity of the crops.

The development of robust detection models depends heavily on the availability of high-quality, annotated datasets. Agricultural settings pose specific obstacles to acquiring such data. While some systems utilize artificial lighting during nighttime operations to maintain visual consistency [151, 104], controlling illumination during standard daytime tasks is typically unfeasible because lighting depends entirely on the time of day, canopy structure, and prevailing weather conditions. Foliage and other structural obstacles can cause sudden illumination shifts, often resulting in under-exposed or over-exposed camera sensors. Furthermore, motion blur is highly common due to the uneven terrain over which robotic platforms navigate. Biological factors also complicate perception. Fruits such as table grapes are self-similar and self-occluding due to their dense, clustered formations.

Combined, these issues make the detection task significantly more complex than

in controlled industrial environments. Due to the biological variability of the crop, the phenomenon of covariate shift, where the statistical properties of training data differ from those encountered during deployment, represents another considerable challenge for the generalization of data-driven algorithms. While methods to detect out-of-distribution performance drops exist [45, 111], the problem of efficiently collecting new labeled data to retrain these models remains open.

Many approaches exist for data generation and augmentation in computer vision. For practical adoption in agriculture, these tools should be accessible enough to be operated by farmers or agronomists, who generally lack specialized computer science skills. Therefore, agronomic data generation pipelines designed to compensate for covariate shifts must be highly automated, allowing end-users to interact with the technology primarily on an agronomic level.

This chapter follows this principle by presenting a system that automatically generates augmented data for object detection, which is a fundamental perception task upon which more complex agronomic operations rely. The only requirement of the proposed method is that a human operator selects a small number of initial sample images collected in the field that are representative of the current season’s crop. Table grapes serve as the primary case study for this approach, as they are subject to severe covariate shifts in appearance across different seasons and are cultivated using training systems (such as the *Tendone* trellis, shown in Figure 4.1) where vine leaves create a dense roof, resulting in highly challenging and uneven illumination.



Figure 4.1. The target operational environment in Aprilia (Italy), characterized by an overhead trellis (*tendone*) system typical for table grape cultivation.

The proposed system utilizes a 3D vineyard simulator developed within the context of the EU project CANOPIES [1], shown in Figure 4.2. Originally built to provide a testing ground for robotic navigation and human-robot interaction, this virtual environment lacks photo-realistic rendering. It represents a typical 3D environment constructed using commonly available graphics assets without requiring highly refined 3D modeling expertise. While the simulator allows for the extraction

of camera sequences and automatic bounding box or segmentation mask generation, training models solely on these purely synthetic fruits does not adequately resolve the covariate shift problem due to the domain gap between simulation and reality.

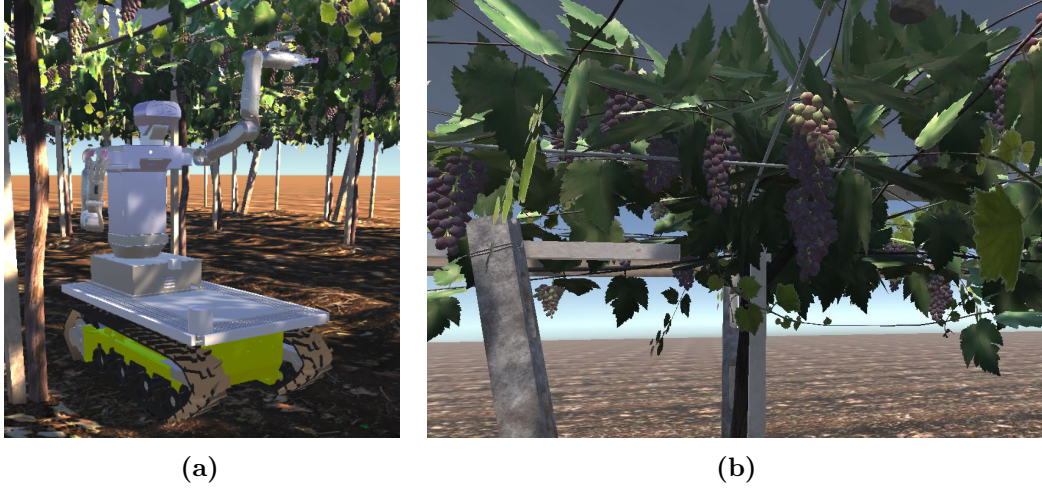


Figure 4.2. (a) The simulated robotic platform, modeled after the Alitrak DCT-300P base and TIAGo++ manipulator, positioned in the virtual vineyard. (b) A synthetic image captured from the perspective of the arm-mounted camera. The relatively uniform textures of these synthetic grapes motivate the need for the proposed cut-and-paste methodology.

To bridge this reality gap, this chapter proposes utilizing real images collected in the field alongside a combination of a semi-supervised object detector and a zero-shot segmentation model to extract real grape instances. Those instances are then pasted onto the synthetic images using various blending strategies to create new, hybrid samples that overcome illumination and seasonal shifts.

The primary contribution detailed in this chapter is the automated generation pipeline, which provides a realistic, scalable solution to the annotation bottleneck and can be easily adapted for other crops.

4.1 Related Work

A comprehensive analysis of the implications of applying robotics and data-driven technologies to agriculture highlights a significant gap between research and practical adoption [125]. Despite decades of advancements in agricultural robotics, widespread integration remains limited [69]. A primary barrier is the unfavorable cost-benefit trade-off associated with current practices for data collection, curation, and utilization. These data-centric processes are fundamental for ensuring the generalization capabilities of machine learning models, yet they often demand prohibitive investments in time and expert labor.

Consequently, label-efficient learning and synthetic data generation have become prominent research areas within the smart agriculture field. An extensive review of the literature identified dozens of studies dedicated to these methods [84]. A common strategy involves generative sampling, where new labeled instances are synthesized

either from a subset of the existing data or entirely from scratch. Generative Adversarial Networks (GANs) [47] are frequently employed for this purpose due to their capacity to produce realistic artificial samples [94]. GANs have been applied across diverse agronomic contexts, including the generation of images of plant seedlings [96], *Arabidopsis* plants [143], soil moisture [52], whitefly pests [71], and the augmentation of datasets for crop and weed segmentation [41]. Despite their success, these networks are notoriously difficult to train, typically requiring highly curated datasets and specialized tuning expertise to avoid instability [94].

As an alternative, direct data synthesis using graphics engines and computer-aided design (CAD) software represents another established strategy to mitigate data scarcity in computer vision [112]. Within the agricultural domain, researchers have leveraged modern simulation engines to automatically extract relevant visual data and generate corresponding ground-truth labels [10]. While there is an abundance of open-source resources for modeling basic natural scenes, developing highly realistic agricultural environments remains resource-intensive and demands advanced 3D modeling skills.

Recognizing the visual limitations of purely synthetic 3D engines, several approaches have attempted to bridge the reality gap by combining CAD-generated samples with GANs, often utilizing architectures like CycleGAN [173] to refine the synthetic textures [54, 9]. Although these combined strategies can yield substantial performance improvements, they inherit the same automation challenges due to the inherent training instabilities of adversarial networks.

Building upon this context, the methodology proposed in this chapter also employs a synthetic environment for sample generation. The approach, however, deliberately avoids the necessity for expert machine learning tuning or specialized 3D graphics engineering. Instead, it relies on simple, highly automatable computer vision algorithms to construct a complete generation pipeline. By blending real fruit instances with simulated backgrounds, the system achieves a level of realism sufficient to deliver consistent performance improvements without the overhead of complex generative models.

4.2 Materials and Methods

The proposed pipeline is designed for the self-supervised generation of realistic agricultural data by merging a simulated environment with real-world textures. The overall architecture of the process, illustrated in Figure 4.3, operates in two primary phases: the automated extraction of real fruit instances and the geometrically consistent blending of these instances into a virtual environment.

4.2.1 Synthetic Data Acquisition

The foundation of the synthetic data generation relies on a Unity-based simulator [109] developed for the CANOPIES Project. This environment replicates a traditional *Tendone* vineyard trellis system, characterized by a three-meter spacing between plants and a dense overhead canopy, similar to the experimental field introduced in Section 2.1.3. The simulated robot (Figure 4.2a) is modeled after the hardware used in field experiments, consisting of a tracked Alitrak DCT-300P mobile base equipped

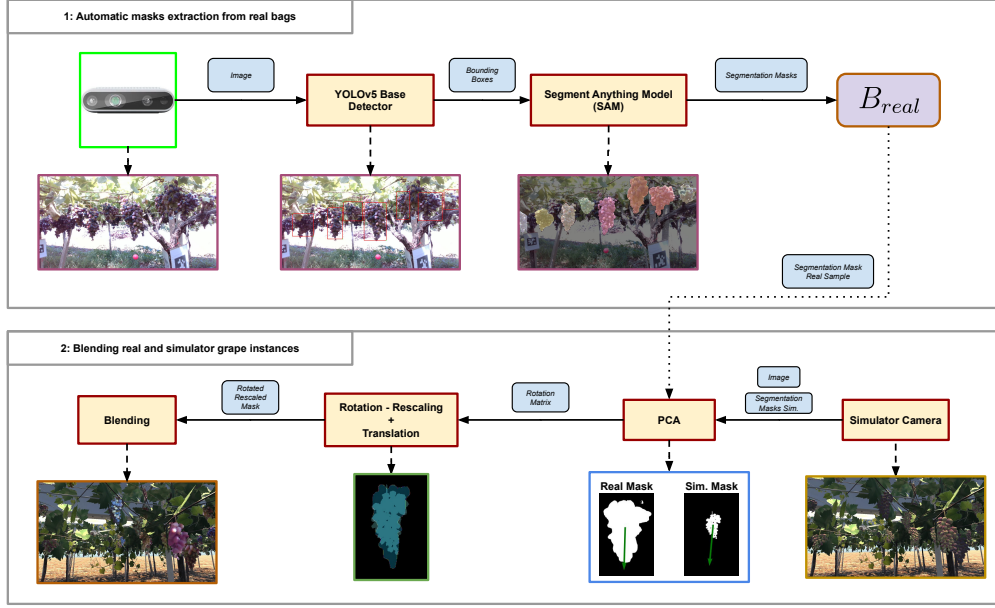


Figure 4.3. Overview of the proposed hybrid data generation pipeline, consisting of two main stages. 1) Automatic mask extraction: a YOLOv5 base detector extracts bounding boxes from real vineyard images, which serve as prompts for the Segment Anything Model (SAM) to generate instance segmentation masks stored in a buffer B_{real} . 2) Blending real and simulated instances: for each synthetic grape instance, a real mask is randomly sampled from B_{real} . Principal component analysis (PCA) is applied to compute the orientation of both masks, allowing the real instance to be rotated, scaled, translated, and ultimately blended over the synthetic target.

with a modified PAL Robotics TIAGo++ humanoid manipulator. A simulated RealSense D435i depth camera is mounted on the end-effector, providing both RGB color images and ground truth segmentation masks with a resolution of 1280×720 pixels.

The simulation environment allows for the programmatic configuration of various agronomic parameters, including the number, size, and density of table grape bunches within the scene. Additionally, the position and intensity of the primary light source, representing the sun, can be set. To ensure visual variability and to make the downstream models robust to camera sensor over-exposure phenomena, the lighting parameters are systematically varied across four equally spaced intensity values, ranging from the minimum to the maximum allowable simulator limits, during data collection. To acquire the synthetic datasets, the simulated robot is navigated along the vineyard rows with the arm-mounted camera directed toward the vine trees, mimicking the typical viewpoint required for harvesting and monitoring operations (Figure 4.2b).

4.2.2 Real-World Instance Extraction

While the simulator provides structural context and free geometric ground truth, the purely synthetic grape textures lack the complex variability found in nature. The

core strategy of this methodology is to extract segmentation masks from real images to isolate the actual grape instances, and then fuse these segmented textures into the simulator. This introduces authentic morphological and textural diversity into the training distribution, which is necessary for training models capable of generalizing to real vineyards.

To avoid the prohibitive costs of manual pixel-level annotation and to provide a fully automated pipeline, the system builds directly upon the weakly supervised methods established in Chapter 3. Specifically, the YOLOv5 [66] base detector trained using the automatically generated pseudo-labels is employed to predict bounding boxes around real grape bunches. These bounding boxes are subsequently utilized as prompts for SAM [74] to automatically extract precise segmentation masks of the real table grapes, as shown in Figure 4.4. The isolated mask and texture data for each extracted instance are then saved into a buffer, denoted as B_{real} , making them available for the subsequent pasting stage.



Figure 4.4. Example of the automated mask extraction process. (a) Initial bounding box predictions generated by the YOLOv5 base detector, which was trained on pseudo-labels. (b) The corresponding instance segmentation masks extracted by the Segment Anything Model (SAM), using the predicted bounding boxes as input prompts.

4.2.3 Context-Aware Cut-and-Paste Blending

The final stage of the methodology involves pasting the extracted real instances from B_{real} into the simulation environment. Rather than employing the random spatial pasting common in standard cut-and-paste augmentations [38], the proposed approach adopts a structured, context-aware approach.

By using the ground-truth spatial coordinates and masks of the synthetic bunches rendered by the simulator, the system ensures a highly realistic positioning of the pasted grapes. For every synthetic bunch present in a simulated frame, a real instance is randomly sampled from B_{real} . To generate a photo-realistic composite and maintain geometric coherence, the real mask must be aligned to match the orientation and scale of the target synthetic bunch.

To compensate for the orientation discrepancy between the real bunch mask M_{real} and the simulated bunch mask M_{sim} , Principal Component Analysis (PCA) is performed on both masks. This extracts the first principal components P_{real} and P_{sim} , which correspond to the longest axis of each respective bunch:

$$\begin{aligned} P_{real} &= PCA(M_{real}) \\ P_{sim} &= PCA(M_{sim}) \end{aligned} \quad (4.1)$$

The real mask is then rotated to align with the simulated mask. The minimum angle of rotation θ_{rot} between the two principal axes is computed, accounting for both possible directional senses of the axis line to prevent an inverted paste:

$$\begin{aligned} \theta_{real} &= \text{atan}_2(P_{real}) \\ \theta_{1,sim} &= \text{atan}_2(P_{sim}) \\ \theta_{2,sim} &= \text{atan}_2(-P_{sim}) \\ \theta_{rot} &= \min(|\theta_{real} - \theta_{1,sim}|, |\theta_{real} - \theta_{2,sim}|) \end{aligned} \quad (4.2)$$

To address the scale difference, the scaling factors f_w and f_h are computed based on the width and height of the real ($real_w$ and $real_h$) and simulated (sim_w and sim_h) segmentation masks:

$$\begin{aligned} f_w &= sim_w/real_w \\ f_h &= sim_h/real_h \\ f &= \max(f_w, f_h) \end{aligned} \quad (4.3)$$

where sim_w and sim_h denote the width and height of the simulated target cluster mask, while $real_w$ and $real_h$ represent the dimensions of the real-world extracted instance mask. The real image instance is then resized using the maximum scaling factor f . Using the maximum factor avoids unnatural texture stretching and ensures that the real texture is large enough to cover the underlying synthetic bunch.

The final step before blending involves clipping the real mask by retaining only the pixel regions that intersect with the underlying synthetic mask. This seamlessly preserves occlusions caused by leaves, branches, or other overlapping grape bunches in the 3D scene. The final composite images thus have a significantly higher degree of realism and structural complexity (Figure 4.5).

4.3 Experiments

The experiments described in this Section demonstrate how combining real and synthetic data through the proposed cut-and-paste technique improves the performance of a detector trained only on automatically generated pseudo-labels [26, 25].

4.3.1 Training Details

The state-of-the-art object detection algorithm chosen for the experiments was YOLOv8 [67]. All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 3090 GPU with 24 GB of VRAM, an AMD Ryzen 9 5950 16-core processor, and 64 GB of RAM. The training runs consist of 100 epochs with a batch size of 8, employing early stopping with a patience parameter of 50 epochs.



Figure 4.5. Example of a hybrid training image generated by the proposed pipeline. Real-world grape instances are geometrically aligned and blended into the synthetic vineyard background using the context-aware pasting method detailed in Section 4.2.3.

A one-cycle learning rate policy [133] was applied, with an initial learning rate of 1×10^{-4} and a final learning rate of 1×10^{-6} . The optimizer utilized was Stochastic Gradient Descent (SGD) with a momentum of 0.937 and a weight decay of 5×10^{-4} .

While the YOLOv8 architecture offers five different scale versions with parameter counts up to 68.2 million, this study focuses exclusively on the "nano" (3.2 million parameters) and "small" (11.2 million parameters) variants. These compact models possess the lowest parameter counts, making them suitable for real-time deployment on limited robotic hardware platforms.

All models were initialized with weights pre-trained on the COCO dataset [89] prior to fine-tuning. The baseline model was trained strictly on the pseudo-labeled dataset using the methodology established in the previous chapter. Conversely, the proposed models were trained on the synthetic and hybrid sets generated via the methodology described in Section 4.2.3. To ensure that any observed performance gains could be attributed only to the introduction of the synthetic data, no additional image augmentation techniques were applied during training.

4.3.2 Evaluation Metrics

The experimental outcomes are evaluated using standard metrics for object detection tasks [107, 89]: Precision, Recall, F1-Score, Intersection over Union (IoU), and mean Average Precision (mAP). Precision measures the proportion of true positive predictions out of all positive predictions made by the model, reflecting its correctness when a detection occurs. Recall represents the ratio of true positive instances to the total number of actual positive instances in the dataset, indicating the model's ability to identify all relevant objects. The F1-Score is the harmonic mean of Precision and Recall, providing a single balanced metric that accounts for both false positives and

false negatives:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4.4)$$

Average Precision (AP) computes the area under the precision-recall curve for a given class. Because mAP is obtained by averaging the individual AP values across all object classes, and this study considers only a single class (table grapes), the AP is mathematically equivalent to the overall mAP. The $mAP_{0.5}$ is defined using an IoU threshold of 0.5, which indicates the model’s performance on relatively easy detections. Additionally, $mAP_{0.5:0.95}$ is computed by averaging the AP across ten distinct IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05, providing a comprehensive evaluation of the model’s localization performance.

4.3.3 Results and discussion

Tables 4.1 and 4.2 present quantitative performance of the YOLOv8 nano and small detectors, respectively, trained using the data generated by the proposed pipeline. Both models are compared across five distinct training splits:

- **Pseudo:** The baseline pseudo-labels dataset established in Chapter 3, consisting of 2400 images with labels automatically generated from video sequences captured in 2021 with an Intel RealSense D435i camera.
- **Synthetic:** A dataset consisting of 1889 images directly captured from the simulator, utilizing the native synthetic ground truth segmentation masks. As described in Section 4.2.1, this set incorporates various illumination conditions: 209 images with low light, 707 with medium light, 559 with high light, and 416 under severe backlighting.
- **Synthetic + Pseudo:** The union of the *Synthetic* and *Pseudo* datasets, totaling 4289 images.
- **Pasted:** A dataset containing the same 1889 images as the *Synthetic* dataset, but with the real-world instances pasted over the simulated bunches using the methodology detailed in Section 4.2.
- **Pasted + Pseudo:** The combination of the *Pasted* and *Pseudo* datasets, also totaling 4289 images.

The testing data originates from a benchmark designed for the multiple object tracking of table grapes [128], which was re-adapted for this object detection evaluation. The *CloseUp* sequences contain 300 frames each, captured in 2023 using an Intel RealSense D435i camera. The *Overview* sequences consist of 100 frames each, captured in 2021 with the same sensor, and feature a more distant perspective similar to the images in the *Pseudo* dataset. The covariate shift between the training set and these test sequences is introduced by differences in the year of capture, illumination conditions, foliage occlusions, and camera distance from the grapes.

The baseline models, trained exclusively on the *Pseudo* set, achieve the highest precision scores for both the small and nano architectures. However, the recall and mAP values for these baseline models are poor across most test sequences.

Table 4.1. Quantitative performance evaluation of the YOLOv8-nano model across four distinct test sequences. The table compares the effect of different training data configurations on Precision, Recall, F1-Score, and mean Average Precision (mAP). The highest value for each metric within a specific test sequence is highlighted in bold.

Sequence	Method	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$	mAP $_{0.5}\uparrow$	mAP $_{0.5:0.95}\uparrow$
CloseUp1	Pseudo	0.793	0.137	0.233	0.454	0.123
	Synthetic	0.431	0.303	0.356	0.350	0.121
	Synthetic + Pseudo	0.696	0.230	0.346	0.432	0.166
	Pasted	0.657	0.493	0.563	0.561	0.237
	Pasted + Pseudo	0.781	0.316	0.450	0.539	0.206
CloseUp2	Pseudo	0.989	0.297	0.457	0.628	0.392
	Synthetic	0.416	0.388	0.401	0.380	0.165
	Synthetic + Pseudo	0.886	0.368	0.520	0.626	0.375
	Pasted	0.609	0.454	0.520	0.548	0.317
	Pasted + Pseudo	0.908	0.462	0.612	0.701	0.433
Overview1	Pseudo	0.950	0.544	0.692	0.741	0.348
	Synthetic	0.097	0.147	0.116	0.055	0.017
	Synthetic + Pseudo	0.883	0.483	0.624	0.649	0.287
	Pasted	0.440	0.425	0.432	0.369	0.146
	Pasted + Pseudo	0.879	0.557	0.681	0.706	0.309
Overview2	Pseudo	0.877	0.297	0.444	0.593	0.301
	Synthetic	0.328	0.266	0.294	0.250	0.082
	Synthetic + Pseudo	0.768	0.449	0.566	0.611	0.287
	Pasted	0.522	0.420	0.465	0.481	0.244
	Pasted + Pseudo	0.758	0.546	0.635	0.667	0.336

Integrating purely synthetic data (the *Synthetic* split) improves recall in certain sequences. This demonstrates that adding simulated images can help the model detect a higher number of instances, proving the underlying potential of simulation to enhance generalizability. The model trained on the *Pasted* set outperforms its purely synthetic counterpart across all metrics. This validates the core hypothesis that blending real instances into a synthetic scenario provides a richer, more effective learning signal than relying on synthetic data alone.

Nevertheless, relying solely on synthetic backgrounds (whether using standard synthetic grapes or pasted real instances) results in a consistent degradation of precision and mAP compared to the baseline. These models generate a high number of false positives because the background environments rendered by the simulator lack complete photo-realism and features overly dense, repetitive foliage. Without exposure to real-world backgrounds during training, the network is frequently led to misclassify complex, real-world background clutter as foreground objects during inference.

By combining the hybrid images with the original real-world data, the model effectively leverages the strengths of both domains. This combined approach improves both recall and mAP values, surpassing the baseline performance. In particular, the inclusion of the *Pseudo* set anchors the model’s understanding of realistic backgrounds, successfully mitigating the high number of false positive detections observed when training on the *Synthetic* and *Pasted* sets. The detector can better distinguish single clusters in cluttered cases and more reliably differentiate the

Table 4.2. Quantitative performance evaluation of the YOLOv8-small model across four distinct test sequences. The table compares the effect of different training data configurations on Precision, Recall, F1-Score, and mean Average Precision (mAP). The highest value for each metric within a specific test sequence is highlighted in bold.

Sequence	Method	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$	$mAP_{0.5} \uparrow$	$mAP_{0.5:0.95} \uparrow$
CloseUp1	Pseudo	0.814	0.215	0.340	0.496	0.186
	Synthetic	0.557	0.391	0.459	0.434	0.133
	Synthetic + Pseudo	0.720	0.330	0.452	0.490	0.192
	Pasted	0.571	0.393	0.465	0.474	0.192
	Pasted + Pseudo	0.777	0.346	0.479	0.527	0.234
CloseUp2	Pseudo	0.965	0.322	0.483	0.647	0.406
	Synthetic	0.547	0.524	0.535	0.531	0.257
	Synthetic + Pseudo	0.863	0.446	0.588	0.672	0.400
	Pasted	0.615	0.557	0.584	0.599	0.324
	Pasted + Pseudo	0.838	0.502	0.628	0.665	0.420
Overview1	Pseudo	0.911	0.582	0.710	0.746	0.377
	Synthetic	0.463	0.308	0.370	0.285	0.084
	Synthetic + Pseudo	0.901	0.585	0.709	0.752	0.367
	Pasted	0.486	0.358	0.416	0.366	0.157
	Pasted + Pseudo	0.845	0.554	0.669	0.667	0.350
Overview2	Pseudo	0.854	0.430	0.572	0.645	0.312
	Synthetic	0.308	0.359	0.331	0.312	0.130
	Synthetic + Pseudo	0.776	0.423	0.547	0.616	0.284
	Pasted	0.457	0.419	0.437	0.377	0.181
	Pasted + Pseudo	0.672	0.544	0.601	0.614	0.298

background from actual grapes.

A qualitative visual representation of the inference differences among the models is provided in Figure 4.6. The model trained on *Pasted + Pseudo* obtains better results primarily due to the successful identification of clusters closer to the camera and can better disambiguate multiple grape clusters when they occlude each other.

4.4 Concluding Remarks

This chapter presented an automated pipeline to improve the performance of object detectors for agricultural applications by generating photo-realistic hybrid data from a simulated environment. By shifting from the surrogate supervision techniques discussed in Chapter 3 to a compositional synthesis approach, the methodology significantly reduces the dependency on exhaustive field data collection. While real fruit instances must still be acquired, capturing diverse environmental contexts, lighting conditions, and viewpoints is achieved by complemented the in-field data with hybrid simulation data, thereby significantly relaxing the physical limitations of the vineyard.

The proposed pipeline automatically extracts grape segmentation masks from real field photos using a combination of a semi-supervised base detector and a zero-shot segmentation model. These authentic real-world instances are subsequently blended onto synthetic images extracted from a virtual 3D vineyard using a context-aware

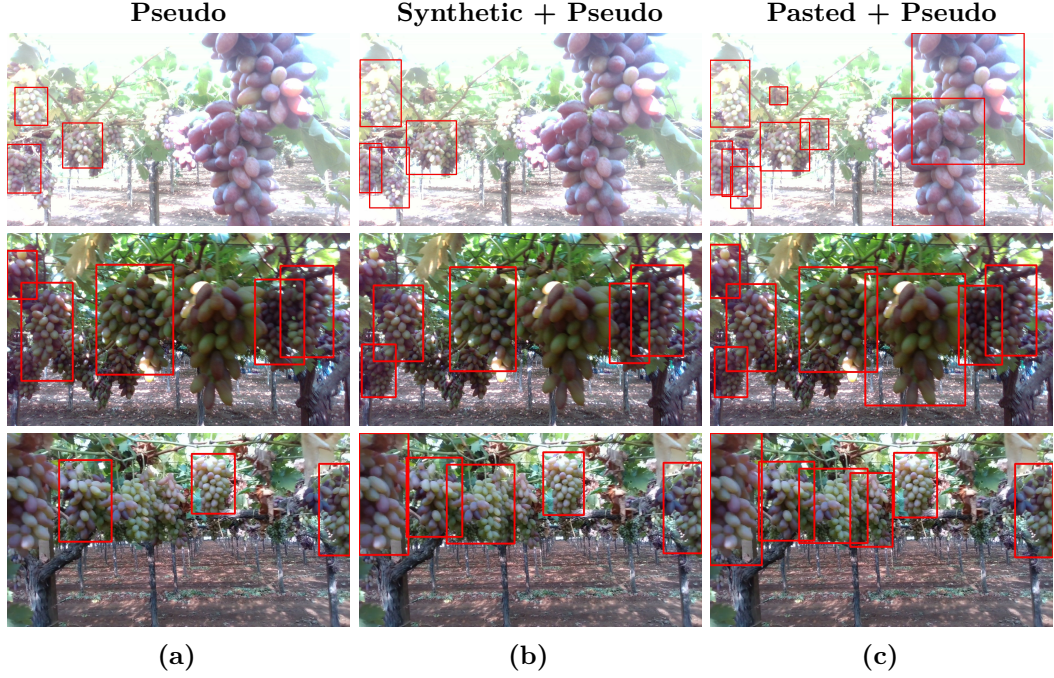


Figure 4.6. Qualitative comparison of YOLOv8 nano inference results across different test sequences. Rows display sample frames from the *CloseUp1* (top), *CloseUp2* (middle), and *Overview2* (bottom) test sequences. Columns denote the respective training datasets: (a) the *Pseudo* baseline, which exhibits low recall and numerous false negatives due to poor generalization; (b) *Synthetic + Pseudo*, which improves upon the baseline but still struggles in highly cluttered areas; and (c) *Pasted + Pseudo*, the proposed model, which demonstrates superior detection capabilities handling occlusions, intense illumination variations, and large foreground clusters.

cut-and-paste approach. To ensure geometric consistency and improve the structural realism of the synthetic composites, the pipeline aligns the real instances with the simulated targets by rotating, rescaling, and translating the extracted masks based on principal component analysis.

Experimental results demonstrate that supplementing an initial, limited dataset with this synthesized data yields considerable improvements over the baseline models. The resulting detector is better equipped to disambiguate highly cluttered instances and exhibits improved generalization capabilities against occlusions, illumination changes, and camera motion blur.

While generating diverse environmental contexts addresses data scarcity for standard fruit detection, comprehensive agricultural perception systems must also identify highly localized surface defects. The detection of anomalies, such as specific diseases or pest infestations, presents a distinct manifestation of the annotation bottleneck because naturally occurring examples are inherently scarce. Addressing this challenge requires a shift from simulating general environmental diversity to developing targeted synthesis strategies. The following chapter explores how to effectively model and synthesize these complex localized anomalies to support robust field perception.

Chapter 5

Synthetic Data Generation for Anomaly Detection

A comprehensive agricultural perception system requires not only the ability to detect and segment general crop structures for automated tasks like robotic harvesting (Figure 5.1a), but also the capacity to identify localized surface defects, such as diseases, pest infestations, and other quality-impacting threats (Figure 5.1b). Simulating the complex textures of these specific anomalies presents a distinct methodological challenge.

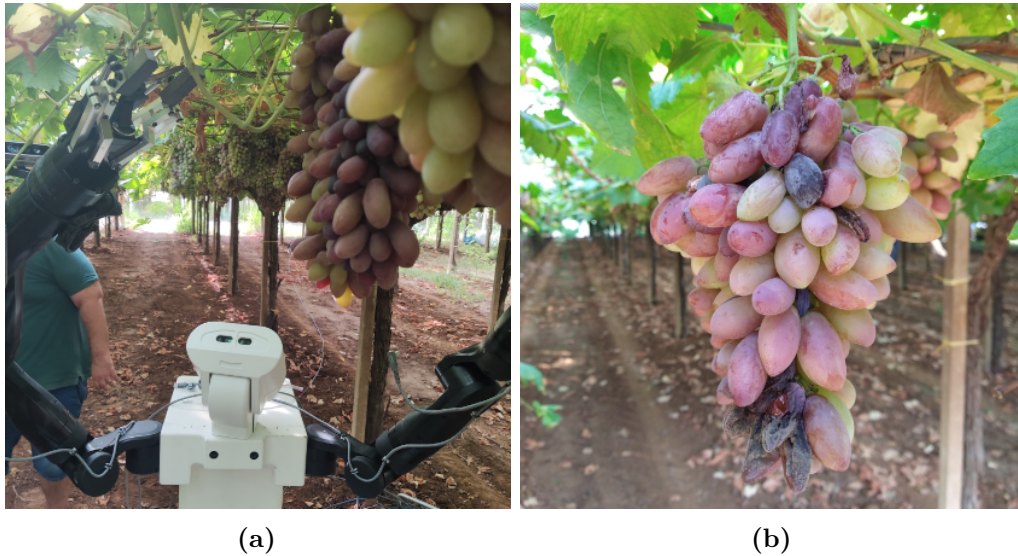


Figure 5.1. (a) A robotic platform executing an automated harvesting operation. (b) Example of localized anomaly. High perceptual accuracy is required to detect such defects, as false negatives can result in harvesting compromised fruits or spreading pathogens. Because naturally occurring examples of specific defects are exceptionally rare, synthetic data generation provides a strategy to address this challenge.

Anomaly detection represents an extreme manifestation of the annotation bottleneck introduced in Chapter 1. Finding sufficient naturally occurring examples of specific defects to train robust supervised classifiers is often impractical, and

the problem is heavily compounded by the persistent covariate shifts inherent to unstructured field environments [138]. For instance, collecting a large, balanced, and diverse dataset of specific diseases (e.g., powdery mildew, botrytis cinerea) is rarely achievable during a single growing season. Consequently, acquiring the data necessary to continuously adapt models to these shifts becomes a prohibitive bottleneck for long-term deployment [3].

For these reasons, early approaches to anomaly detection in agriculture often relied on unsupervised or semi-supervised modeling of the normal condition of the fruit, for which there is plenty of data. In these frameworks, any sample exhibiting a distribution significantly different from the healthy training set is flagged as an anomaly. For instance, [137] trained an auto-encoder on patches of healthy grapevines, utilizing a reconstruction loss threshold to distinguish normal from anomalous patches and generating spatial heatmaps to locate defects. Subsequent works improved upon this by employing Variational Auto-Encoders [97] or applying dimensionality reduction techniques like PCA to hyperspectral images [93].

While single-class training gracefully bypasses the scarcity of defective samples, it is generally unsuited for identifying specific types of anomalies and is highly susceptible to false alarms triggered by complex background clutter and illumination noise common in real orchards. For this reason, the direct approach of training supervised models to specifically recognize target anomalies remains highly relevant. Some studies have attempted supervised detection by training Convolutional Neural Networks (CNNs) on cropped images of healthy and damaged fruits [14]. However, these datasets are typically small and lack the covariate shifts encountered in real-world applications. To scale the data volume, other researchers have utilized web-scraped datasets to train YOLO-based anomaly detectors [152]. Although these datasets are larger, they do not model the actual distribution of any specific agricultural field.

Given these fundamental limitations, significant research has been devoted to data augmentation and synthetic data generation. Specific augmentation schemes, such as channel randomization [24], randomly permute color channels to generate unnatural color compositions during a self-supervised pretext task, encouraging representations that are sensitive to color irregularities. These methods are useful when anomalies are primarily expressed through pigmentation changes and serve as complementary techniques to standard geometric transformations (rotations, cropping) or random noise injection [102, 86]. However, simple image manipulations are insufficient when the target samples are exceptionally rare. An alternative strategy involves training generative algorithms to synthesize entirely new samples. For example, [103] applied CycleGAN [173] to detect anomalies in mandarins. While generative models can be effective, they are notoriously difficult to train, highly sensitive to hyperparameters, and converging on the correct anomalous distribution is rarely a controllable or easily automated process.

This chapter proposes a novel compositional synthesis approach. By combining classical computer vision techniques with foundational models, the proposed pipeline generates highly realistic synthetic anomalous samples from a minimal set of real vineyard images. The algorithm is designed to operate with minimal manual intervention, requiring only a basic initial separation of normal and anomalous training examples by an agronomist. Once provided with this initial categorization,

the method autonomously synthesizes realistic anomalous data for specific field conditions.

The main contributions of this chapter are:

- A novel, texture-focused, semi-automatic algorithm for synthetic data generation that leverages foundation models for instance extraction and classical edge detection for structural analysis.
- A curated dataset of normal and anomalous table grape images collected from the experimental vineyard.

The proposed method is validated through experiments using a baseline CNN classifier, demonstrating significant improvements in performance metrics, particularly under severely constrained labeling budgets.

5.1 Materials and Methods

5.1.1 Data Collection and Labelling

As introduced in Chapter 2, the experimental activities were conducted in a commercial vineyard located in southern Lazio, Italy. The focus of this study was on the Black Pizzutello variety to maintain consistency with the experiments in previous chapters. However, the proposed methodology is designed to be adaptable and can be extended to other varieties without loss of generality.

Field images were collected using a smartphone camera during the 2021 growing season to build the dataset for the synthetic generation pipeline. An expert agronomist performed the initial categorization, identifying instances based on the presence of visible mold, rot, insect damage, pest infestations, and other structural or quality-impacting issues. From an initial set of 88 images of healthy grape bunches and 41 images containing confirmed anomalies, 512x512 pixel patches were extracted to isolate the regions of interest. This extraction resulted in a final dataset consisting of 529 healthy grape patches and 166 patches exhibiting distinct anomalies. For all subsequent experiments and model evaluations, a 3-fold cross-validation strategy was employed to ensure robust statistical validation across the dataset splits.

5.1.2 Synthetic Data Generation Procedure

The core of the proposed methodology is the synthetic data generation procedure, which builds upon the limited set of real anomaly samples described in the previous section. The pipeline assumes that a preliminary detection module, such as the bounding box framework discussed in Chapter 3, has already isolated and cropped the grape bunches from the wider field images. It is further assumed that square patches are extracted from these crops and subjected to a basic binary categorization (normal versus anomalous) by a domain expert. One of the key strengths of the proposed approach is its robustness; even with some label noise in this initial manual categorization, the synthesis method generates realistic anomalous samples with a low rate of false positives. Figure 5.2 visually summarizes the overall architecture of the generation method, while the specific functional components are detailed below.

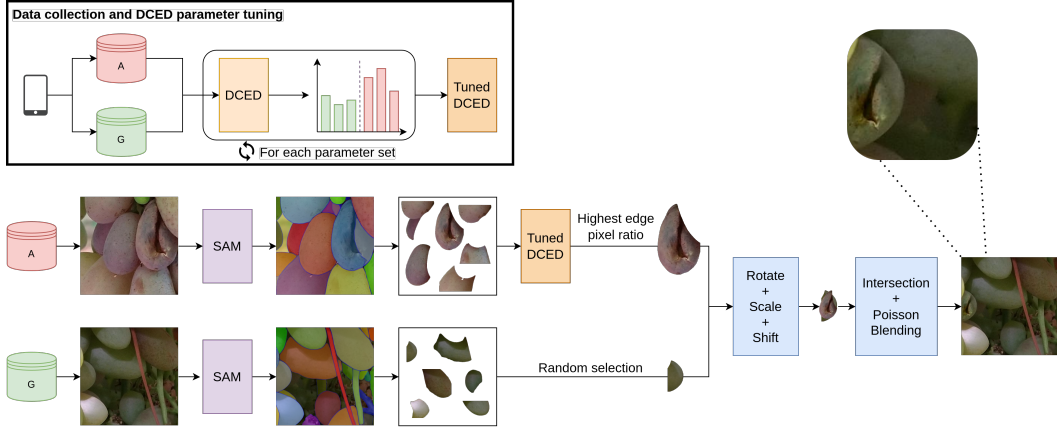


Figure 5.2. Architecture of the proposed compositional synthesis pipeline. The upper section illustrates the preliminary categorization of real data and the tuning of the Dual-Canny Edge Detection (DCED) parameters. The lower section details the automated synthesis process. The Segment Anything Model (SAM) extracts instance masks from a paired set of real training images. The anomalous berry is isolated using the highest internal edge pixel ratio determined by the DCED, while the target healthy berry is randomly selected. Following geometric alignment (rotation, scaling, and translation), the intersection of the masks is computed, and Poisson blending seamlessly integrates the textures to output a realistic synthetic anomaly.

Dual-Canny Edge Detection

A key observation for automatically isolating anomalous berries relies on texture analysis. Healthy table grapes typically exhibit a relatively smooth surface, whereas anomalous berries are characterized by distinct, irregular patterns depending on the specific disease or defect. Based on this observation, the system leverages classical computer vision to measure texture roughness via edge detection filtering, specifically utilizing the Canny Edge Detection (CED) algorithm [17]. Figure 5.3 illustrates the application of CED for this specific purpose.

The performance of CED is highly dependent on the selection of specific parameters, particularly the hysteresis thresholds, denoted hereafter as th_{max} and th_{min} . Following an initial Gaussian smoothing step with a $K \times K$ kernel to suppress sensor noise, the CED algorithm computes the image gradient to identify potential edges. A double-thresholding process then refines these candidates: pixels exceeding th_{max} are strictly classified as strong edges, pixels below th_{min} are discarded as noise, and pixels falling between the two thresholds are retained only if they are spatially connected to a strong edge. This strategy guarantees edge continuity while filtering out minor intensity fluctuations.

Although edge-based statistics could theoretically be derived from a single CED pass, the external boundaries of the berries are inevitably present in both normal and anomalous samples, confounding the textural analysis. To isolate the internal surface irregularities, the pipeline introduces a Dual-Canny Edge Detection (DCED) filter. This technique involves executing the CED algorithm twice with distinct threshold configurations, followed by a subtraction operation. The first execution

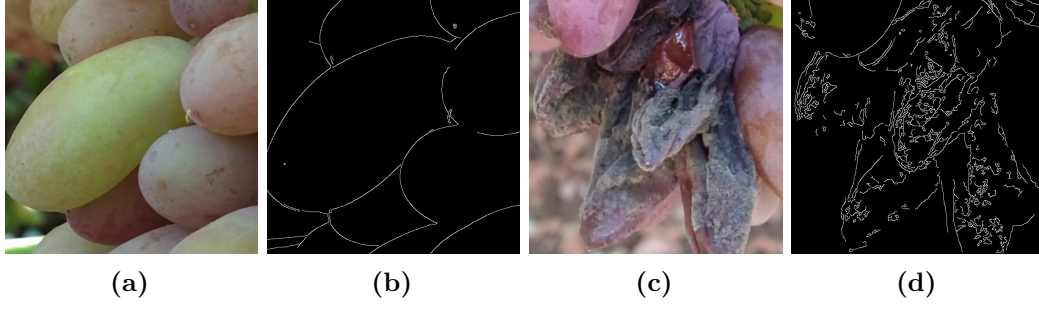


Figure 5.3. Textural analysis of table grape patches using the Canny edge detector. A patch containing healthy berries (a) yields an edge map (b) consisting almost entirely of external contours, reflecting the smooth surface of the fruit. In contrast, a patch containing anomalous berries (c) produces a highly complex edge map (d), demonstrating that specific defects manifest as distinct textural roughness that can be captured via structural edge features.

(wide-CED) applies a wide hysteresis region with lower threshold values (wth_{min} and wth_{max}), capturing a high density of edge pixels. The second execution (narrow-CED) applies stricter, higher threshold values (nth_{min} and nth_{max}), isolating only the most prominent edges, typically the external berry contours. By subtracting the narrow edge map from the wide edge map, the resulting image predominantly isolates the pixels associated with internal surface defects, effectively filtering out the shared boundary features. Figure 5.4 demonstrates the qualitative differences between standard CED passes and the proposed DCED filter.

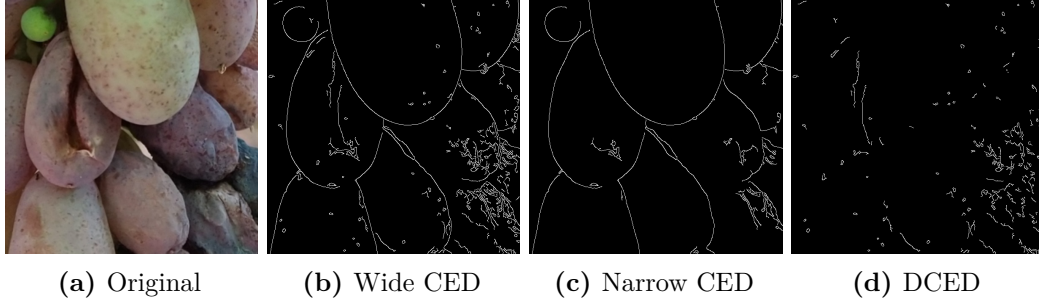


Figure 5.4. Dual-Canny Edge Detection (DCED) filter applied to isolate surface defects. For a given anomalous sample (a), two separate edge maps are generated using a wide hysteresis threshold (b) and a stricter, narrow threshold (c). The final DCED map (d) represents the difference between the two edge spaces. This subtraction effectively isolates the edges corresponding to the anomalous internal texture while suppressing the shared external boundaries of the berries.

To optimize the DCED parameters for accurately separating healthy and anomalous patches, the system iterates over various combinations of the Gaussian blur kernel size K and the four DCED thresholds (wth_{min} , wth_{max} , nth_{min} , nth_{max}). The search space is constrained by considering only multiples of 25 within the range $[0, 250]$ and by enforcing the logical constraints: $wth_{max} > wth_{min}$, $nth_{min} \geq wth_{min}$, $nth_{max} > nth_{min}$, and $nth_{max} > wth_{max}$.

For each combination of parameters, the system analyzes the resulting DCED

edge pixel counts across the training set to identify the optimal separation threshold, utilizing balanced accuracy as the primary evaluation criterion. These parameter combinations and their corresponding separation thresholds are subsequently evaluated on the validation set. Patches with an edge pixel count exceeding the threshold are classified as anomalous, while those falling below are classified as normal. The optimal DCED parameter set is selected based on the highest overall validation performance and is subsequently integrated into the synthetic sample generation phase. Furthermore, these optimized parameters and the edge pixel count threshold establish a deterministic baseline classifier for the experimental comparisons detailed in Section 5.2.

Berry Segmentation and Merging

The final phase of the compositional synthesis technique involves segmenting damaged berries from the identified anomalous patches and seamlessly blending them onto healthy berries within normal patches. Most importantly, this process must preserve the natural geometry, scale, and orientation of the destination berry to maintain realism. Figure 5.5 provides a schematic overview of this procedure.

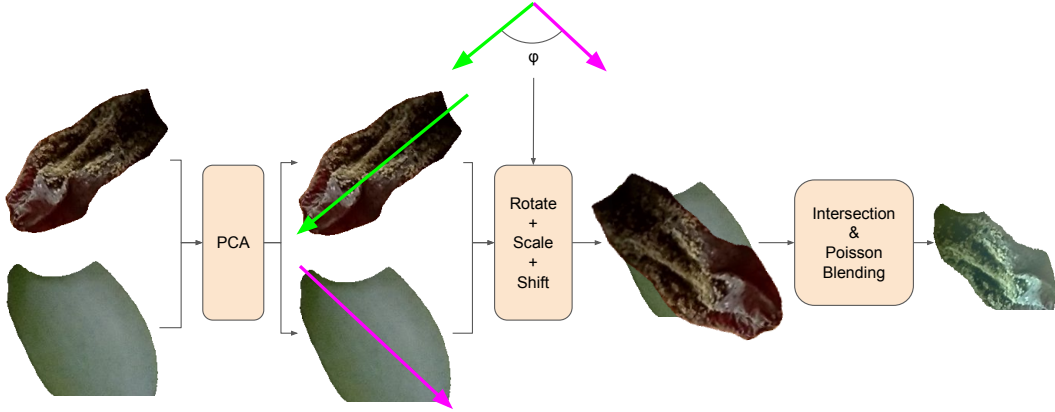


Figure 5.5. Geometric alignment and blending process for compositional synthesis. Principal Component Analysis (PCA) is applied to the source and destination masks to extract their primary axes of variation (indicated by the green and pink arrows). The rotation angle ϕ and scale factor are computed to align the anomalous instance with the target healthy berry. Following spatial translation, the intersection of the two masks is extracted, and Poisson blending is applied to seamlessly fuse the textures, yielding a realistic synthetic composite.

To bypass the need for manual pixel-level annotation, the pipeline leverages SAM’s [74] Automatic Mask Generator mode, which extracts all identifiable object masks from the patch without requiring external prompts like bounding boxes or points. To refine the output, the generated masks are filtered to retain only the top 50% based on pixel area and strictly those exceeding the mean mask area of the patch. While being straightforward, this filtering heuristic proves highly effective in practice because the anomalies of interest are typically localized on individual berries, which constitute the largest connected components in the patch. Although some background elements, such as leaves or structural supports, may

occasionally be captured by the mask generator, empirical observations confirm that such occurrences are infrequent and exert a negligible impact on the overall distribution of the synthetic dataset.

As SAM indiscriminately segments both anomalous and healthy berries alongside potential background artifacts, the pipeline automates the specific selection of the anomalous instance by integrating the previously tuned DCED filter. Given a set $\mathcal{S} = \{s_0, \dots, s_n\}$ of n segmented elements extracted from a source anomalous patch $\mathcal{P}_S$, the algorithm analyzes the textural profile of each mask and selects the mask s_i with the highest ratio of internal edge pixels as the primary anomaly target. This targeted selection process is formalized in Algorithm 1. Conversely, the destination mask d_j is randomly selected from the set of m masks $\mathcal{D} = \{d_0, \dots, d_m\}$ extracted by SAM from the target healthy patch $\mathcal{P}_D$.

Algorithm 1 Select Anomalous Instances

Require: masked images of the *berries*, number of berries to select n , wide canny edge detector parameters $wide_params$, narrow canny edge detector parameters $narrow_params$

- 1: $edge_ratios \leftarrow$ Empty list
- 2: **for** $berry \in berries$ **do**
- 3: $berry \leftarrow \text{GaussianBlur}(berry)$
- 4: $n_wide \leftarrow \text{CountNonZero}(\text{CannyEdgeDetector}(wide_params))$
- 5: $n_narrow \leftarrow \text{CountNonZero}(\text{CannyEdgeDetector}(narrow_params))$
- 6: $n_berry \leftarrow \text{CountNonZero}(berry)$
- 7: $edge_ratios \leftarrow (n_wide - n_narrow)/n_berry \cup edge_ratios$
- 8: **end for**
- 9: $edge_ratios \leftarrow \text{SortDescending}(edge_ratios)$
- 10: **return** $edge_ratios[:n]$

To generate a realistic composite image, the segmented anomalous source berry must be seamlessly aligned and blended with the normal target berry. While naive data augmentation strategies often paste instances at random spatial coordinates within the target image [44], this pipeline employs a structurally aware approach, mirroring the geometric alignment techniques introduced in Chapter 4.

Specifically, the longitudinal orientation of both berries is computed using Principal Component Analysis (PCA). The PCA algorithm is applied to the spatial coordinates of the source mask s_i and the destination mask d_j to extract their respective principal axes, denoted as Z_{s_i} and Z_{d_j} . The source anomaly s_i is subsequently rotated by an angle ϕ_{ij} and scaled by a factor γ_{ij} to match the destination geometry, defined mathematically as:

$$\gamma_{ij} = \frac{\text{Area}(d_j)}{\text{Area}(s_i)} \quad (5.1)$$

$$\phi_{ij} = \arccos \left(\frac{Z_{s_i} \cdot Z_{d_j}}{|Z_{s_i}| |Z_{d_j}|} \right) \quad (5.2)$$

Following rotation and scaling, the transformed anomalous berry s'_i , is spatially translated onto the centroid of the normal target berry. To ensure the pasted

texture does not spill beyond the natural boundaries of the target fruit, the intersection of their respective segmentation masks is computed and applied. Finally, to eliminate harsh boundaries and harmonize the illumination between the source texture and the target environment, Poisson image blending [116] is performed. The complete automated pipeline for generating these synthetic datasets is summarized in Algorithm 2.

Algorithm 2 Synthetic Data Generation

Require: Set of normal images N , set of anomalous images A , number of synthetic samples per anomaly n_{syn}

```

1: DCED  $\leftarrow$  TuneDCED( $N, A$ )
2:  $SyntheticData \leftarrow \emptyset$ 
3: for  $img\_bad \in A$  do
4:    $img\_good \leftarrow \text{Sample}(N)$ 
5:    $bad\_berries \leftarrow \text{filter}(\text{SAM}(img\_bad))$ 
6:    $bad\_berries \leftarrow \text{SelectAnomalousInstances}(bad\_berries, DCED, n_{syn})$ 
7:    $good\_berries \leftarrow \text{filter}(\text{SAM}(img\_good))$ 
8:   for  $i = 1$  to  $n_{syn}$  do
9:      $good\_berry \leftarrow \text{Sample}(good\_berries)$ 
10:     $bad\_berry \leftarrow bad\_berries[i]$ 
11:     $bad\_berry \leftarrow \text{Rotate}(bad\_berry, good\_berry)$ 
12:     $bad\_berry \leftarrow \text{Scale}(bad\_berry, good\_berry)$ 
13:     $synthetic\_img \leftarrow \text{PoissonBlending}(bad\_berry, good\_berry)$ 
14:     $SyntheticData \leftarrow SyntheticData \cup synthetic\_img$ 
15:   end for
16: end for
17: return  $SyntheticData$ 

```

Figure 5.6 provides visual examples of the synthetic generation process, illustrating both successful composites and edge-case failures. While the pipeline generally achieves seamless, realistic blending, extreme scale discrepancies between source and target berries can occasionally result in unnatural texture stretching. Furthermore, rare segmentation errors may lead to background regions being incorrectly targeted. These edge cases underscore the importance of geometric constraints and indicate potential areas for future refinement, such as stricter scale-matching heuristics and enhanced background suppression.

5.2 Experiments and Results

The experimental evaluation was conducted using a 3-fold cross-validation strategy over the dataset, with all reported metrics representing the mean results across the three independent folds.

5.2.1 Baselines

To establish a rigorous standard for evaluating the synthetic data generation pipeline, two distinct baselines were defined. The first baseline represents a purely classical

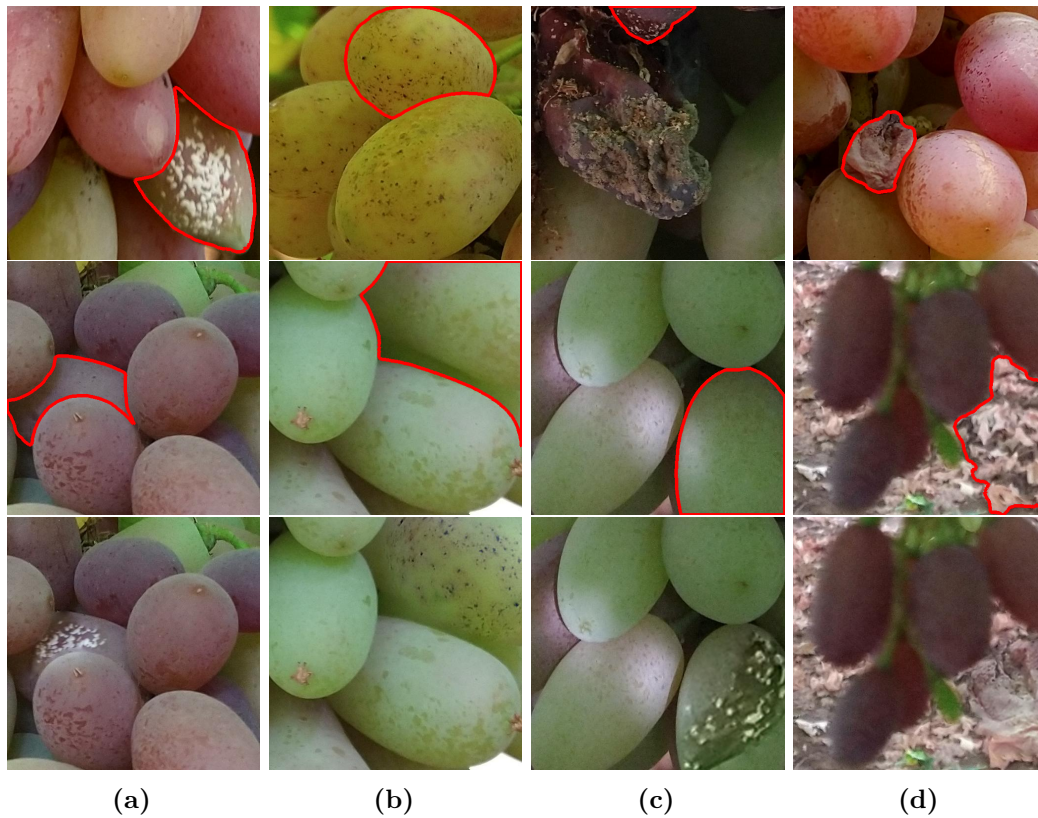


Figure 5.6. Qualitative examples of the compositional synthesis process. Segmented anomalous berries (first row) are integrated into healthy target images (second row) to produce synthetic anomalous composites (third row), with the respective source and destination boundaries outlined in red. Columns (a) and (b) illustrate highly successful blending operations that seamlessly maintain a realistic appearance. Conversely, column (c) demonstrates a failure mode where extreme scale discrepancies between the source and target berries yield unnatural texture distortion. Finally, column (d) illustrates an edge case where the automated mask selection incorrectly targets a background region, resulting in an invalid synthetic anomaly.

computer vision approach, applying the Dual-Canny Edge Detection (DCED) filter directly to the image patches for classification. Utilizing the optimal DCED parameters and threshold, the filter was applied to the validation split. The results, detailed in Table 5.1, provide a strong deterministic baseline for comparison.

The second baseline is a deep neural network classifier. This model was constructed by fine-tuning a ResNet18 [56] backbone, with its final fully connected layer modified to accommodate the binary classification task. A comprehensive hyperparameter sweep was conducted to determine the optimal training configuration. The model was trained for 100 epochs using the Adam optimizer, with an initial learning rate of $1e-5$. To facilitate fine convergence, the learning rate was reduced to $1e-6$ for the final 10 epochs. The training process utilized a batch size of 32 and incorporated standard data augmentations, including random horizontal flips and color jitter. Model weights were saved at the epoch achieving the highest balanced accuracy. As shown in Table 5.1, the ResNet18 baseline significantly

Table 5.1. Performance comparison of the baseline models.

Baseline	Balanced Acc.	F1-Score	Precision	Recall
DCED & threshold	76.99	60.70	48.99	80.06
ResNet18	94.99	92.92	93.99	91.88

outperformed the deterministic DCED approach, and it consequently serves as the primary benchmark for all subsequent evaluations.

It is critical to contextualize these performance metrics within the agronomic domain. In anomaly detection, the cost of a false negative is substantially higher than that of a false positive. A false negative implies failing to detect a diseased or infested fruit, which can lead to the unchecked spread of pathogens throughout the orchard, resulting in severe crop loss and economic damage. Because the ResNet18 baseline already achieves high performance, demonstrating statistically significant improvements upon it using synthetic data is particularly challenging, making it a robust benchmark for validating the proposed synthesis pipeline.

5.2.2 Experiments

The following experiments investigate the extent to which augmenting the dataset with synthetic samples generated by the proposed pipeline can improve the performance of the baseline DNN classifier. Because the original training set already contains a limited number of authentic anomalous samples, these experiments aim to measure whether this compositional synthesis provides a beneficial training signal.

For each fold in the cross-validation process, the optimal DCED parameters were first extracted and subsequently fed into the synthetic sample generation module. The pipeline was configured to generate exactly one synthetic image for every real anomalous sample present in the training set, applying the procedure described in Section 5.1.2. No synthetic samples were generated for the healthy class. To evaluate the impact of this generated data, varying proportions (10%, 25%, 50%, and 100%) of the synthetic samples were randomly sampled and added to the original training dataset. Furthermore, to explicitly examine the covariate shift between real and synthesized textures, a parallel set of substitution experiments was conducted. In these trials, 10%, 25%, 50%, and 100% of the real anomalous data were removed and replaced entirely with synthetic counterparts.

All models in this phase were trained using the exact hyperparameters and augmentation strategies established for the ResNet18 baseline, and were not optimized for the augmented dataset.

As shown in Table 5.2, the average results demonstrate that adding synthetic data via the proposed method consistently yields improvements in both balanced accuracy and F1-score. The maximum gain in balanced accuracy was achieved at the 10% and 50% addition thresholds, reaching 95.35. The highest F1-score of 93.30 was obtained by adding 25% of the synthetic data. However, a marginal performance plateau or slight decrease was observed when 100% of the synthetic samples were added. This behavior aligns with findings in other data synthesis studies, suggesting that the optimal ratio of synthetic to real data must be carefully calibrated to avoid

overwhelming the natural distribution [101]. Conversely, substituting real samples with synthetic ones consistently caused a degradation in performance. This critical finding highlights that while compositional synthesis provides a powerful method to augment constrained labeling budgets, it is not a complete replacement for capturing authentic field variability.

Table 5.2. Performance impact of augmenting or substituting the training set with single-berry synthetic anomaly samples.

Configuration	Balanced Accuracy	F1-Score	Precision	Recall
ResNet18 Baseline	94.99	92.90	93.99	91.88
Addition 10%	95.35	93.06	93.36	92.78
Addition 25%	95.24	93.30	94.48	92.18
Addition 50%	95.35	93.06	93.42	92.77
Addition 100%	94.74	92.40	93.40	91.56
Substitution 10%	72.33	61.24	96.88	45.23
Substitution 25%	87.69	84.88	95.68	76.53
Substitution 50%	91.84	88.85	91.93	86.14
Substitution 100%	55.81	25.93	56.14	17.48

As the generation of synthetic images is fully automated, there remains a marginal risk of the foundation model inadvertently selecting background elements (e.g., leaves or branches) as the source or target for anomaly generation. While pasting background elements onto healthy berries can still qualify as a synthetic anomaly, the goal is to avoid scenarios where background elements from the source are pasted exclusively onto the background of the target. To mitigate the likelihood of generating images without any valid synthetic anomalies, a secondary strategy was tested where three distinct berries were sampled and pasted to create a single synthetic image.

As presented in Table 5.3, this multi-instance pasting approach led to slight but consistent improvements in both balanced accuracy and F1-score compared to the single-berry method. The peak performance was achieved by augmenting the training set with an additional 25% of synthetic images. However, consistent with the previous findings, substituting real samples with these multi-berry synthetic images proved detrimental.

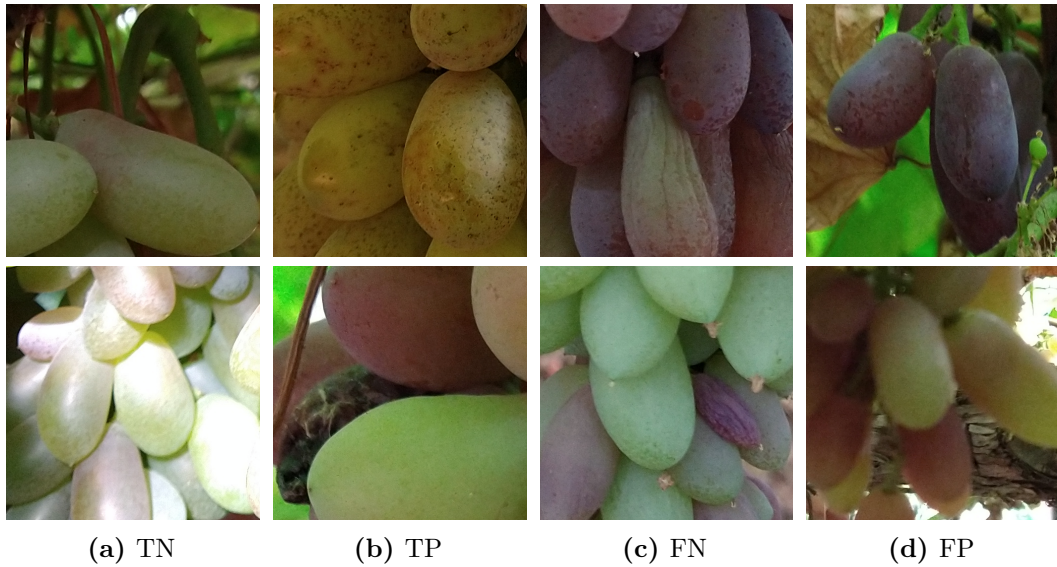
Figure 5.7 provides qualitative classification examples from the best-performing model (ResNet18 augmented with 25% multi-berry synthetic data). These visual results highlight the model’s robustness to varying illumination and background clutter (true negatives) and its success in identifying complex defects (true positives). However, the presence of false negatives reveals ongoing challenges in detecting anomalies that exhibit only subtle visual deviations from healthy grapes, indicating potential areas for future refinement in the synthetic generation process.

5.3 Concluding Remarks

This chapter addressed the extreme data scarcity inherent to agricultural anomaly detection by introducing a novel compositional synthesis methodology. To overcome

Table 5.3. Performance impact of augmenting or substituting the training set with multi-berry (three instances) synthetic anomaly samples.

Configuration	Balanced Accuracy	F1-Score	Precision	Recall
ResNet18 Baseline	94.99	92.90	93.99	91.88
Addition 10%	95.45	93.33	93.91	92.78
Addition 25%	95.75	93.70	94.16	93.39
Addition 50%	95.53	93.62	94.54	92.77
Addition 100%	95.02	93.25	95.11	91.56
Substitution 10%	73.03	62.74	97.36	46.43
Substitution 25%	87.30	84.22	94.96	75.93
Substitution 50%	91.82	89.36	93.87	85.53
Substitution 100%	58.69	32.34	52.98	24.76

**Figure 5.7.** Qualitative classification results from the best-performing model on the validation set (ResNet18 augmented with 25% multi-instance synthetic data). Column (a) presents true negatives, demonstrating the model’s robustness to complex background elements and extreme illumination variances. Column (b) displays true positives, confirming the successful detection of authentic field anomalies. Conversely, column (c) highlights false negatives, revealing the ongoing challenge of identifying subtle or early-stage defects that visually mimic healthy textures. Finally, column (d) presents false positives, where healthy grapes are misclassified, an error likely caused by out-of-focus capture or misleading background textures.

the limitations of the annotation bottleneck, the proposed approach combined the zero-shot capabilities of foundational models with the structural priors of classical computer vision. This integration enabled the automatic segmentation, identification, and seamless blending of anomalous textures onto healthy grape images, generating a realistic and diverse set of synthetic training samples from a minimal initial dataset.

Experimental evaluations confirmed that incorporating this synthesized data into the training pipeline improved the performance of a baseline deep neural network, yielding gains in both balanced accuracy and F1-score. Furthermore, the experiments

highlighted that while compositional synthesis provides a highly effective mechanism to augment constrained labeling budgets, it functions best as a complement to, rather than a complete substitute for, authentic field variability.

Future research could further refine the instance selection heuristics, explore advanced geometric blending techniques, and integrate domain-specific agronomic constraints to maximize the quality of the generated data. Ultimately, this synthetic generation framework offers a scalable, label-efficient solution for deploying robust anomaly detection systems in complex agricultural environments.

While the proposed methodology addresses the identification of localized structural and textural defects, a fully autonomous robotic harvesting system must also evaluate the internal chemical ripeness of the fruit to make informed operational decisions. Moving beyond surface-level geometry requires inferring chemical properties that are not strictly visible. Consequently, the following chapter investigates how agricultural perception systems can be extended to estimate internal quality metrics, specifically Soluble Solid Content, using standard camera sensors.

Chapter 6

Estimating SSC with Simple RGB Sensors

Executing a successful automated harvesting operation requires a robotic system to evaluate the internal chemical ripeness of the fruit before making a final harvesting decision. Measuring internal quality metrics, specifically Soluble Solid Content (SSC), introduces a distinct manifestation of the annotation bottleneck. While geometric properties like bounding boxes or segmentation masks can be visually annotated from standard images, establishing ground truth for internal chemical properties requires specialized instrumentation, such as contact spectrometers or analog refractometers. This physical requirement limits the volume of data that can be efficiently collected, complicating the training of robust deep learning models.

While hyperspectral imagery (HSI) shows strong potential for this task because its specific wavelength responses correlate with the absorption patterns of chemical properties, deploying these sensors in unstructured agricultural settings presents practical challenges. Hyperspectral cameras are typically expensive, slow to acquire images, sensitive to illumination conditions, and complex to integrate into mobile platforms (Figure 6.1). Therefore, developing a passive SSC estimation system using affordable RGB cameras remains an open and relevant research objective. Table 6.1 summarizes the fundamental differences between HSI and standard RGB imaging devices for in-field applications.

This chapter addresses this gap by validating the feasibility of RGB-based estimation and proposing high-performance, label-efficient algorithmic solutions tailored for practical deployment in agricultural settings. The approach is directly modeled on standard harvesting procedures, where an expert agronomist first confirms a target SSC using a refractometer, and subsequent workers harvest individual bunches based on correlated visual cues. While SSC and color are not strictly causally linked, they share latent ripening factors including climate, weather, training practices, and nutrients availability [62]. These shared factors allow human experts to visually approximate ripeness. Therefore, the main objective is to replicate and scale this expert capability using data-driven methods paired with accessible RGB sensors.

The primary contributions detailed in this chapter are:

- A multi-season, multi-sensor dataset of in-field grape images labeled with both SSC (°Brix) and agronomic color categories. This represents one of the first



Figure 6.1. The robotic harvesting platform deployed in the experimental vineyard. Accurate quality assessment relies on the wrist-mounted and head-mounted RGB sensors to perform real-time Soluble Solid Content (SSC) estimation, simulating the visual checks performed by human workers.

table grape datasets combining field imagery directly with °Brix ground truth labeling to enable quality estimation studies under realistic field conditions.

- An analysis of the visual-to-SSC correlation, including a user study that establishes a baseline for human-level performance in estimating ripeness from imagery alone.
- The design and validation of two distinct algorithms for real-time, calibration-free operation: a lightweight, histogram-based method for resource-constrained robotic systems, and a high-performance multi-task deep neural network capable of robust estimation under limited labeling budgets.

6.1 Hyperspectral and RGB Approaches

To contextualize this study, the following subsections provide a brief review of the relevant literature regarding both hyperspectral and RGB applications for fruit quality assessment. For a broader overview of fruit ripeness estimation methods, particularly for grapes, we refer the reader to [110, 146]

6.1.1 Hyperspectral applications to agricultural robotics

Robots have been increasingly used for monitoring and quality estimation across different crops and are finding industrial applications in high-value cultivations like wine grapes [156, 117], strawberries [160, 159], and blueberries [155]. Several studies have explicitly explored the quality estimation of fruits using hyperspectral sensors.

In [49], an HSI device is mounted on an Autonomous Ground Vehicle (AGV) to measure the Total Soluble Solids (TSS), a synonym of SSC, of Tempranillo wine grapes. The AGV moves along the vineyard row acquiring spectral data

Table 6.1. Comparison of Hyperspectral (HSI) and RGB Imaging devices for in-field use.

Feature	Hyperspectral Imaging (HSI)	Standard RGB Imaging
Spectral Resolution	High (hundreds of narrow bands, usually 5–10 nm)	Low (3 broad bands: Red, Green, Blue)
Spatial Dimension	Low	High
Information Captured	Detailed spectral signatures, correlated with chemical composition	Surface color, texture, and shape
Equipment Cost	€10000 - €200000+	€10 - €1000+
Operational Environment	Laboratory or highly controlled settings	In-field or indoors alike
Acquisition Time	High (seconds to minutes, depending on illumination)	Low (few milliseconds)
Computational Load	High (due to the processing 3D hypercubes)	Low (suitable for real-time processing on embedded systems)
Portability & Size	Often bulky and requires careful calibration	Compact, and easily integrated into robotic platforms

in a push-broom fashion. To achieve color constancy, the authors perform an indirect calibration for exposure and an illuminant calibration every 5 meters using a white reference. Although this setup demonstrates the feasibility of on-the-go HSI estimation, achieving a Root Mean Square Error (RMSE) as low as 1.27 °Brix, the authors note the difficulty of maintaining the steady motion required by push-broom cameras on an AGV. The required stop-and-go operations for calibration and the lengthy post-processing time of approximately 5 hours make this approach ill-suited for real-time robotic harvesting. Utilizing a simple RGB camera would eliminate both the vibration sensitivity and the complex calibration requirements.

A similar approach utilizing a trailing cart instead of an AGV was proposed by [154] to monitor raspberries. The cart mounts two HSI devices to capture both VNIR (400–896 nm) and SWIR (896–2506 nm) ranges. The setup requires an overarching arm with a black background screen behind the plants and a suspended white reference. While this study successfully demonstrated the physical setup, it did not estimate physical characteristics from the hyperspectral data.

Another highly successful robotic platform called Ladybug [142] is equipped with an HSI device (VNIR 390–887 nm) designed for low horticultural crops. The robot is a straddle model where illumination constancy is maintained through a stroboscopic light to avoid interfering with other optical sensors. Calibration pads are placed both in the field and on the robot at the periphery of the camera’s field of view. While straddle systems offer one of the most promising strategies for deploying HSI in the field, calibration challenges persist, and this overarching design is entirely impractical for taller crops like table grape overhead trellises.

In preparation for robotic use, the authors of [68] conducted thorough experimentation using a VIS-SWIR (400–2500 nm) contact spectrometer to estimate SSC in four wine grape varieties. Contact probe spectrometers offer output similar to HSI imaging devices but are designed for precise point sampling. These devices provide an internal illumination source and must be carried manually. Their cost remains significantly higher than RGB cameras, and their integration into a robotic platform is complicated by the need for delicate physical contact with the fruit. The authors analyzed the contribution of different wavelengths and applied various algorithms including Partial Least Square Regression (PLSR), Random Forest (RF), Convolutional Neural Networks (CNN), and Support Vector Machines (SVM). Across the four grape varieties, they achieved a Root Mean Square Error (RMSE) ranging from 1.70 to 2.20 °Brix.

In a follow-up study [141], the authors introduced the in-field estimation of SSC using a snapshot HSI camera (VNIR 400–1000 nm). A primary contribution of their work is a method to address illuminant estimation without relying on a reference tile before each acquisition, proposing instead to learn the illuminant normalization through an auto-encoder. The ground truth for the SSC estimation was collected with an analog refractometer by sampling the °Brix of a single berry per bunch. Using a neural network to estimate the illuminant distribution shares conceptual similarities with our work by simplifying the acquisition setup. However, it sacrifices the absolute metric accuracy provided by a physical white reference and still requires periodic recalibration using a highly reliable spectrometer. Furthermore, utilizing a snapshot HSI camera considerably reduces spatial resolution compared to push-broom setups. The authors explored four wine grape varieties, reporting RMSE values ranging from 1.66 to 2.20 °Brix.

While these approaches demonstrate the notable successes of HSI in agricultural robotics (typically achieving RMSEs between 1.27 and 2.20 °Brix), they also exhibit several key limitations, including high sensor costs, operational complexity, lengthy processing times, and restricted adaptability to specific crops where overarching machinery is impractical, such as table grapes. In contrast, the methodology proposed in this chapter aims to achieve comparable performance without relying on hyperspectral data, operating in real-time, and being directly suited for edge deployment.

6.1.2 RGB applications to fruit quality assessment

Given the limitations of HSI illuminant calibration and device costs, numerous authors have proposed alternative approaches based on standard RGB sensors. In [12], researchers explored the use of color imaging combined with advanced numerical techniques such as interval numbers and neural network regressors to estimate grape maturity indices including TSS, titratable acidity, and pH. The authors utilized a stereo camera to acquire RGB images under natural daylight, converting them to the CIELAB color space. Their study focused on Tempranillo, a black grape cultivar, which facilitates a clearer correlation between color and SSC. While conceptually similar to our approach, their reported performance errors are considerably higher, yielding an approximate RMSE of 4.63 °Brix.

In [158], an analytical study was conducted on Kyoho grape berries using an

industrial RGB camera. The authors proposed the use of specific feature indices computed from standard RGB channels as inputs for Partial Least Squares Regression (PLSR) and Multiple Linear Regression (MLR). They obtained an impressively low cross-validation error of 0.78 °Brix (RMSE) on a dataset with a standard deviation of 1.37°. Although this study was performed under highly controlled laboratory conditions, it successfully demonstrates the underlying effectiveness of data-driven estimations utilizing simple RGB sensors. Building upon these findings, our work demonstrates that sufficiently low estimation errors can also be achieved in unstructured field environments to perform automatic harvesting reliably.

A related approach is presented in [18], which proposes a non-destructive estimation technique based on an industrial RGB camera using random forest predictors trained on color histograms. The histogram was computed in the CIELAB color space by clustering colors and selecting those with the highest intrinsic variance. While this study uniquely targeted white grapes, the estimation problem was framed strictly as a qualitative color scale measurement rather than an explicit SSC regression. The classification task was performed on a 5-level quality scale, and no precise °Brix measurement was included in the labeling process. The experimental setup involved 400 bunches divided between Italia and Victoria varieties, captured under artificial illumination. Image features were computed after background subtraction and meticulous white balancing using color calibration references. While this study cannot be directly compared to ours quantitatively, it proves that even within white grape varieties, careful modeling can capture the latent relationship between color and quality.

This growing body of research confirms both the scientific and practical demand for RGB-based internal crop characteristic estimation. However, existing methods typically either fail to match the performance levels of HSI imaging (e.g., 4.63 °Brix RMSE in field conditions) or struggle to transition effectively from controlled laboratory environments (where they can achieve 0.78 °Brix RMSE) to the complex, uncontrolled lighting of an actual orchard.

6.2 Materials and Methods

This section details the data acquisition and labeling processes, followed by the two algorithmic approaches developed for SSC estimation. Figure 6.2 provides a visual overview of the complete experimental workflow, from data collection to final evaluation.

6.2.1 Dataset

Data collection

The data for these experiments were collected in the commercial vineyard in southern Lazio, Italy, introduced in Chapter 2. Focusing on the Black Pizzutello variety, static images were collected across two summer seasons under realistic working conditions. To simulate the covariate shifts that occur between initial field data collection by an agronomist and eventual deployment on robotic hardware, three distinct camera devices were utilized.

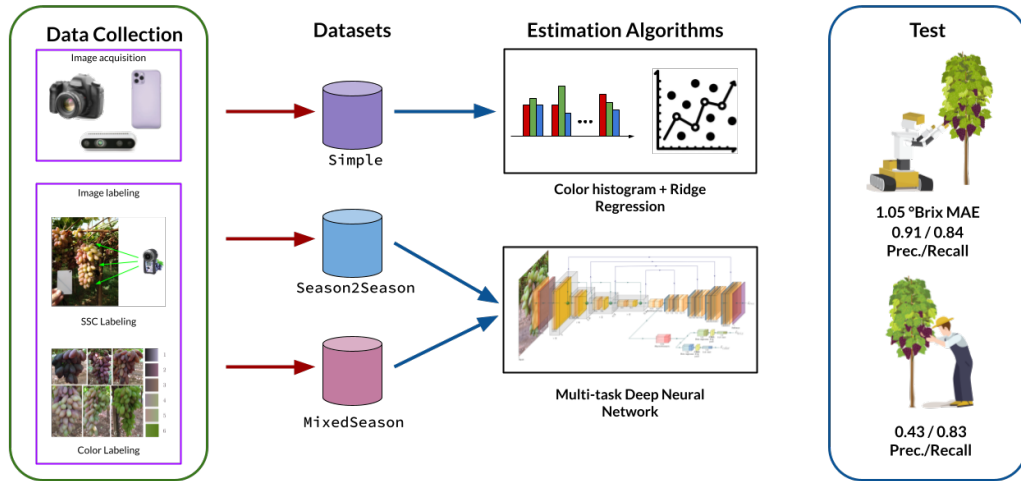


Figure 6.2. An overview of the experimental workflow. The process moves from left to right: (1) In-field Data Collection using multiple RGB sensors and subsequent labeling for SSC and color quality. (2) Organization of the data into three distinct Datasets to evaluate performance under different conditions. (3) Development and validation of two estimation algorithms: a lightweight histogram-based model and a high-performance multi-task deep neural network. (4) The Test panel summarizes the key results, outlining the best model’s SSC estimation error (MAE) and harvesting decision performance (Precision/Recall) against the human expert baseline.

Images were acquired using two different smartphones, a DSLR reflex camera, and an Intel RealSense D435i. The smartphone images were captured with a Moto G8 Plus and a Realme GT Master Edition, respectively at 3000×4000 and 3448×4624 pixel resolutions in JPG and raw formats. The Reflex images were captured in raw format at a 3900×2613 pixel resolution, while the D435i provided RGB images at a lower 1280×720 pixel resolution. When available, raw images were corrected for white balance using a neutral gray card, although images without a gray card were still used for comparison. Figure 6.3 provides examples of images collected with the same smartphone in both standard JPG and white-balanced raw formats.

All images were manually labeled with bounding boxes around the grape bunches



Figure 6.3. Examples of paired standard JPG and raw (white-balanced) images collected with the same smartphone device.

and subsequently cropped. The resulting crops vary in size depending on the distance from the camera and the number of targets in the frame. The combination of capture season, sensor type, and gray card presence creates distinct data distributions, summarized in Table 6.2 as atomic splits. The variations between them introduce the specific covariate shifts studied in our experiments.

Table 6.2. Description of the atomic splits according to the device, image format, gray card presence, and the possibility of having expert ground truth from white-balanced images.

Id	Year	Smartphone				Reflex		D435i			Num images	Num samples	Color GT
		Type	Jpg	Raw	Card	Raw	Card	RGB	IR	Card			
A	2021	motog8	yes	no	no	no	-	no	no	-	100	100	no
B	2021	motog8	no	yes	yes	yes	yes	no	no	-	38	38	yes
C	2021	motog8	yes	yes	yes	yes	yes	no	no	-	62	62	yes
D	2021	motog8	yes	yes	yes	yes	yes	no	no	-	50	50	yes
E	2022	realmi	yes	yes	no	no	-	yes	no	no	1	2	no
F	2022	realmi	yes	yes	yes	no	-	yes	no	yes	6	22	yes
G	2022	realmi	yes	yes	no	no	-	yes	no	yes	1	2	no
H	2022	realmi	yes	yes	yes	no	-	yes	no	yes	4	19	yes
I	2022	motog8	yes	yes	yes	no	-	yes	yes	yes	8	51	yes

Labeling protocol

All images were labeled for both SSC and color quality. For SSC, an Atago Hikari PAL2 digital refractometer was used. This device is capable of measuring berry sugar content in °Brix through surface contact without damaging the berry, operating within a range of 9° to 26°. Typically, an SSC of 16° to 18° is considered sufficient for harvesting depending on the specific grape variety. To accurately replicate the visual estimation performed by human experts, we adopted a labeling procedure mirroring the SSC sampling performed by agronomists. As illustrated in Figure 6.4, because sugar concentration varies across the bunch due to gravity and individual berry development, the SSC was sampled three times from the top, middle, and bottom of the bunch, and the results were averaged. For consistency, the refractometer was placed on the side of the berry in each measurement.

For color, a categorical classification was required. In collaboration with experienced agronomists, a 6-level color quality chart was defined based on standard production use cases. The scale ranges from 1 (highest quality) to 6 (lowest quality). Values of 1 and 2 represent premium produce, 3 is considered acceptable for certain markets, and values from 4 to 6 are deemed unripe. This index accounts for both the overall color saturation and its distribution across the individual berries and the entire bunch. Examples of this color variability are shown in Figure 6.5.

Experimental Dataset Splits

The overall dataset is constructed from the atomic splits detailed in Table 6.2. These minimal, homogeneous groups were combined to create specific experimental datasets designed to evaluate model performance under varying degrees of domain shift. The

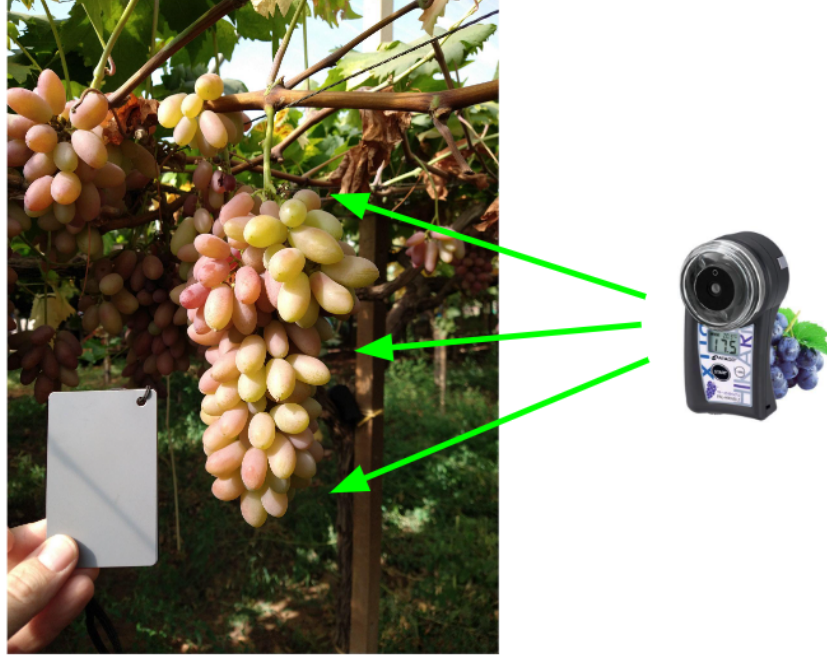


Figure 6.4. Example of an image collected with a gray card for white balance correction. An agronomist performs in-field measurements by sampling the SSC of berries from the top, middle, and lower sections of the bunch using a contact refractometer, subsequently averaging the three values to establish the ground truth.

combined datasets are distinguished by their image processing type, denoted by the subscript **STD** for standard JPG images or **WB** for white-balanced images derived from raw captures. A summary of these configurations is presented in Table 6.3.

Table 6.3. Summary of the experimental dataset splits used for model training and evaluation.

Name	Train set components	Test set components	Description
Simple_{STD}	C+D		Smartphone JPG images (2021)
Simple_{WB}	C+D		Smartphone white-balanced images (2021)
S21ToS22Phone	A+C+D	F+H+I	Smartphone JPG images (2021–2022)
S21ToS22Phone_{WB}	B+C+D	F+H+I	Smartphone white-balanced images (2021–2022)
S21ToS22All_{WB}	B+C+D	F+H+I	Smartphone & DSLR white-balanced images (2021–2022)
MixedSeason_{STD}	A+C+D+H+I	F	All JPG images
MixedSeason_{WB}	B+C+D+H+I	F	All white balanced images

The datasets are grouped by their analytical purpose:

- The **Simple** splits are small, single-season (2021) datasets used for initial exploratory assessments of the models. They do not feature a predefined train/test division.
- The **S21ToS22Phone** splits restrict the sensor type to smartphones and are



Figure 6.5. Examples of the six color levels utilized for categorical quality classification. While a clear distinction between neighboring categories can be challenging for an untrained eye, agronomists classify the bunches by evaluating both the general color shades and the density of highly colored berries. Labeling validation tests indicated consistent classification despite some inherent boundary uncertainty.

designed to evaluate cross-seasonal robustness by training on data from the 2021 season and testing on data from 2022.

- The **MixedSeason** splits represent the most challenging deployment scenario. They are designed to simulate a real-world robotic application by training on smartphone and DSLR reflex images (mimicking initial data collection by an agronomist) and testing exclusively on D435i images (mimicking the final robotic platform sensors). This setup directly tests the models' resilience to hardware-induced covariate shifts.

To provide a quantitative baseline for the experiments, Table 6.4 summarizes the key statistical properties of these datasets. The table details the size of each training and testing split along with the natural variability of the SSC labels, measured by the mean, standard deviation, and Mean Absolute Deviation (MAD). These statistics serve as reference for evaluating the performance of the proposed estimation models.

6.2.2 SSC Estimation Models

This section outlines the two approaches developed for SSC estimation.

Table 6.4. Statistical properties of the experimental dataset splits.

Name	Split	# samples	SSC mean	SSC std	SSC MAD	Description
Simple _{STD}	-	112	17.19	2.43	1.99	Smartphone JPG images (2021)
Simple _{WB}	-	112	17.19	2.43	1.99	Smartphone white-balanced images (2021)
S21ToS22Phone _{STD}	tr	212	16.23	2.52	2.02	Smartphone JPG images (2021)
	te	92	18.91	2.16	1.71	Smartphone JPG images (2022)
S21ToS22Phone _{WB}	tr	150	17.54	2.46	2.04	Smartphone white-balanced images (2021)
	te	92	18.91	2.16	1.71	Smartphone white-balanced images (2022)
S21ToS22All _{WB}	tr	300	17.54	2.46	2.04	Smartphone & DSLR white-balanced images (2021)
	te	92	18.91	2.16	1.71	Smartphone & DSLR white-balanced images (2022)
MixedSeason _{STD}	tr	352	17.67	2.75	2.33	All JPG images, across all seasons and devices
	te	44	15.96	1.68	1.36	The test set contains images with SSC content close to the threshold (16° Brix).
MixedSeason _{WB}	tr	440	18.27	1.98	1.96	All white-balanced images, across all seasons and devices.
	te	44	15.96	1.36	1.36	The test set contains images with SSC content close to the threshold (16° Brix).

Estimation with Histogram Features

The primary use case for these estimators involves deployment on robotic platforms, embedded edge devices, or smartphones, which operate with strictly limited memory and computational power. Consequently, the first approach relies on classical handcrafted features, specifically color histograms. Because the visual estimation task relies fundamentally on capturing color intensity and its spatial distribution, RGB channel histograms serve as the foundational building blocks for this model.

To capture color distribution differences, each cropped bunch image is divided into a grid of rectangular regions, and a color histogram is computed for each region individually. A key observation is that even with tight bounding boxes generated by an object detector, the rectangular regions in the corners frequently contain background foliage due to the natural, tapered shape of the grape bunches. While deploying a secondary deep neural network for background subtraction could isolate the fruit, it would exponentially increase the computational cost.

To address this efficiently, a straightforward heuristic was developed to discard grid regions likely to contain background pixels. Based on the empirical observation that table grape bunches are generally wider in the middle and narrower at the extremities, the algorithm retains only a cross-shaped selection of grid regions, as illustrated in Figure 6.6. The exact dimensions of this pattern are treated as hyperparameters to be optimized during training. Three distinct configurations are

defined in Equation 6.1: "none" (retaining all regions), "fat" (discarding only the four corner regions), and "thin" (discarding corners and additional bottom-edge regions).

These three configurations are programmatically defined by the selection of horizontal (n_{bin_x}) and vertical (n_{bin_y}) grid divisions:

$$\begin{aligned}
 none &: [0, n_{bin_x} - 1, 0, n_{bin_y} - 1] \\
 fat &: \left[\left\lfloor \frac{n_{bin_x}}{4} \right\rfloor, \left\lfloor \frac{3n_{bin_x}}{4} \right\rfloor, \left\lfloor \frac{n_{bin_y}}{4} \right\rfloor, \left\lfloor \frac{3n_{bin_y}}{4} \right\rfloor \right] \\
 thin &: \left[\left\lfloor \frac{n_{bin_x}}{4} \right\rfloor, \left\lfloor \frac{3n_{bin_x}}{4} \right\rfloor, \left\lfloor \frac{n_{bin_y}}{4} \right\rfloor, \left\lfloor \frac{n_{bin_y}}{2} \right\rfloor \right]
 \end{aligned} \tag{6.1}$$

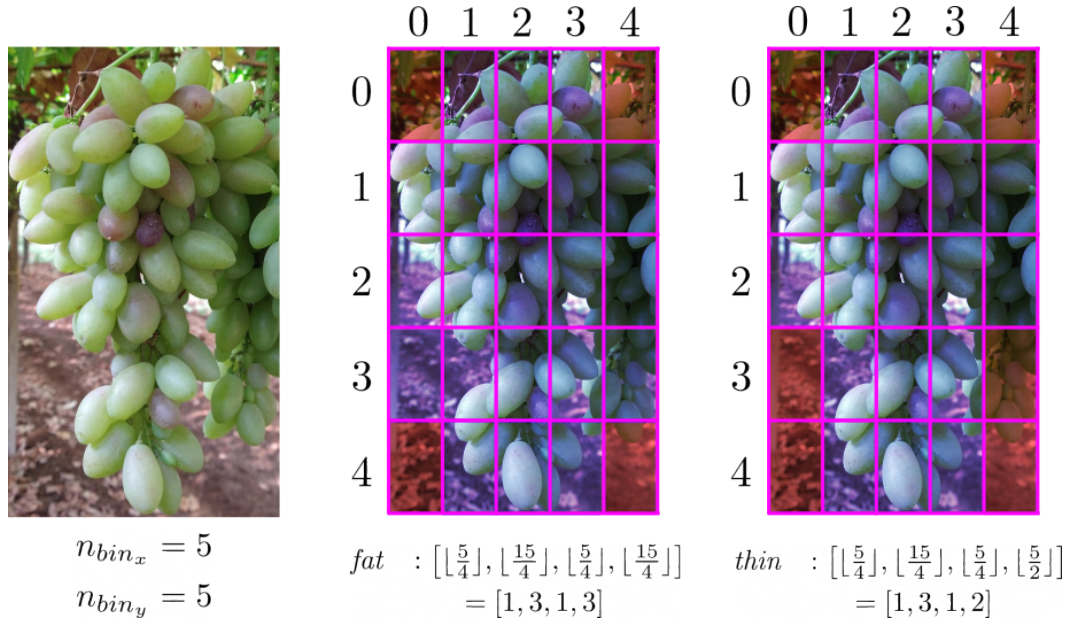


Figure 6.6. An illustration of the cross-shaped feature selection heuristic for histogram-based estimation. An input image is divided into a grid of n_{bin_x} by n_{bin_y} bins (a 5×5 grid is shown). To efficiently remove background pixels without requiring complex segmentation, two structural patterns are tested: a "fat" pattern retaining a wider central cross-section, and a "thin" pattern narrower at the bottom, mirroring the typical shape of a bunch. The indices of the retained bins are determined using Equation 6.1. For the 5×5 example shown, this results in the index ranges $[1, 3, 1, 3]$ for the fat pattern and $[1, 3, 1, 2]$ for the thin pattern.

Averaging the color values within each retained cell and concatenating them produces a final feature vector of length $3 \times n_{bin_x} \times n_{bin_y}$. The experimental validation also explored the effects of initial image resolution and the conversion to Hue-Saturation-Value (HSV) or CIELAB color spaces. Given the small number of training images, Ridge Regression was selected as the estimation algorithm. It provides an interpretable regression output while incorporating L2 regularization to prevent overfitting. While more complex models such as Support Vector Machines or Random Forests could potentially yield higher peak performance, the primary

goal of this specific method is to explore the viability and characteristics of the lightweight histogram features.

Estimation with Multi-Task Learning DNN Architecture

The SSC and color distribution can be viewed as properties conditioned on the broader environmental and agronomic parameters of the field. While both color and SSC depend heavily on the internal chemical state of the berries, this state is itself influenced by macroscopic factors such as local climate, training practices, and soil nutrients. Because the overall visual appearance of the bunches is shaped by these same factors, a neural network can be designed to learn the latent representations linking appearance to internal chemistry.

The architecture of the proposed multi-task deep neural network, illustrated in Figure 6.7, reflects this rationale. To distill a robust feature vector, an Auto-Encoder (AE) architecture is employed rather than relying solely on a standard feature extraction backbone. The encoder component learns to capture common appearance variations that may be correlated with the internal chemical state of the grapes. These shared, compressed features are subsequently leveraged by two separate prediction heads to estimate the agronomic color class and the SSC value simultaneously. In a data-scarce scenario, incorporating the auxiliary reconstruction loss of the auto-encoder forces the network to focus heavily on appearance details, thereby strengthening the back-propagation signal and reducing the likelihood of overfitting to the limited SSC labels.

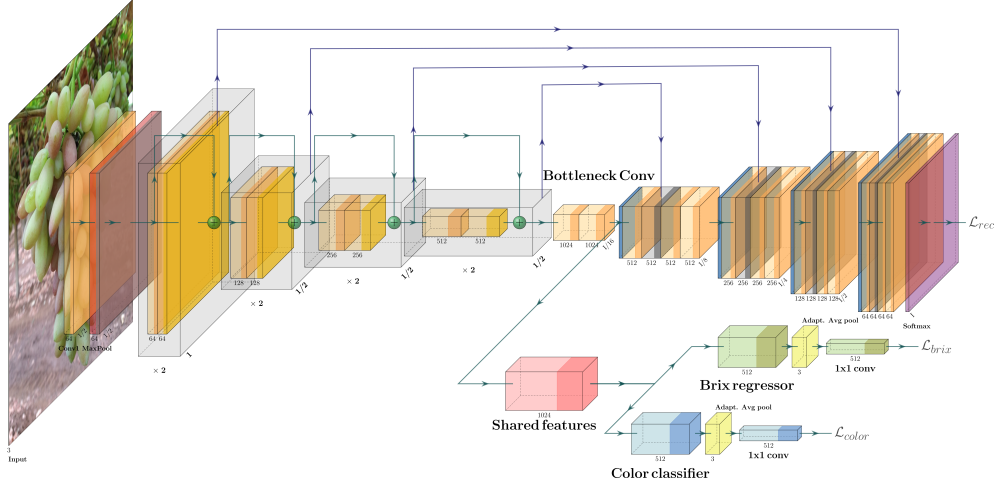


Figure 6.7. The multi-task network architecture utilized for joint SSC and color estimation. The model is based on a U-Net [124] with a ResNet18 [56] backbone acting as an auto-encoder to learn a compact visual representation, driven by the reconstruction loss ($\mathcal{L}_{rec}$). Features from the Bottleneck Conv layer are routed to a Shared features block. The network then branches into two parallel heads: a Brix regressor for continuous SSC estimation ($\mathcal{L}_{brix}$) and a Color classifier for discrete color quality grading ($\mathcal{L}_{color}$). (Diagram generated with PlotNeuralNet [61])

The auto-encoder utilizes a U-Net [124] structure built upon a ResNet18 [56]

backbone. The two prediction heads share a common bottleneck layer comprising a CNN with 512 output channels, ensuring the core visual representation is strictly shared between the two tasks. Standard U-Net architectures are generally unsuitable for simultaneous reconstruction and regression problems because the dense skip connections allow the network to learn a simple identity function, bypassing the bottleneck. To prevent this, the architecture explicitly extracts the regression features directly from the bottleneck output prior to the upsampling decoder stages. From this bottleneck, a convolutional block extracts the features fed to the task-specific heads. Both heads share an identical structure consisting of three CNN blocks that progressively reduce the spatial feature size, followed by adaptive average pooling [171] and a 1×1 convolution [88] for the final estimation. For the color head, the outputs are logits for the multi-class classification problem, while the SSC head outputs a continuous scalar estimation.

The overall network is optimized using a compound loss function that combines the pixel-wise Mean Squared Error (MSE) reconstruction loss ($\mathcal{L}_{rec}$) of the auto-encoder, a Mean Absolute Error (MAE) loss for the continuous SSC estimation ($\mathcal{L}_{brix}$), and a Cross-Entropy loss for the categorical color classes ($\mathcal{L}_{color}$):

$$\mathcal{L}_{MT} = \mathcal{L}_{rec} + \alpha \mathcal{L}_{color} + \beta \mathcal{L}_{brix} \quad (6.2)$$

where α and β are weighting factors used to balance the distinct contributions of the classification and regression losses. These weights were systematically optimized during cross-validation alongside the structural hyperparameters.

To properly evaluate the multi-task network, a custom scoring metric termed the Brix-Color-Hinge (BCH) metric was developed. This metric is designed to balance raw statistical accuracy with practical agronomic requirements. It integrates the continuous regression error for the °Brix estimation (using Mean Absolute Error, $\mathcal{L}_{brix}$) and the classification accuracy for the color grading (using Balanced Accuracy, $\mathcal{A}_{color}$). Because these metrics exist on different scales, a normalization factor η is applied. Furthermore, the metric introduces strict hinge-like penalties for predictions that fall below predefined acceptable operational quality thresholds (t_{brix} and t_{color}). The final BCH score $\mathcal{S}$ to be minimized is formulated as:

$$\mathcal{S} = \eta(1 - \mathcal{A}_{color}) + \mathcal{L}_{brix} + \mathcal{L}_{CH} + \mathcal{L}_{BH}, \quad (6.3)$$

while the hinge penalties for °Brix ($\mathcal{L}_{BH}$) and color ($\mathcal{L}_{CH}$) are defined as:

$$\mathcal{L}_{BH} = \max(0, \mathcal{L}_{brix} - t_{brix})^\rho, \quad (6.4)$$

$$\mathcal{L}_{CH} = \max(0, -\eta(\mathcal{A}_{color} - t_{color}))^\rho, \quad (6.5)$$

where ρ dictates the degree of the penalty and is set to 2 in all our experiments. The threshold t_{brix} is set to 2.5 °Brix (comparable to the standard deviation of the training data), and t_{color} is set to 0.80 (reflecting a typical acceptable human accuracy rate of 80% in this context [139]). The normalization factor is empirically derived as $\eta = t_{brix}/t_{color}$. This formulation ensures that the optimized model prioritizes meeting minimum quality standards before fine-tuning absolute accuracy, making it practically viable for automated harvesting decisions.

6.3 Experiments

This section describes the experimental validation of the proposed algorithmic solutions for SSC estimation. To provide a meaningful perspective on the performance of these data-driven models, an initial set of experiments was conducted to establish a human performance baseline. Consequently, this section begins by detailing the methodology and results of the human evaluation (Section 6.3.1), followed by the experimental results for the lightweight histogram feature model (Section 6.3.2) and the multi-task DNN (6.3.3).

The algorithmic solutions were developed and initially trained locally on a laptop equipped with an AMD Ryzen 9 processor and an NVIDIA GeForce RTX 3080 GPU, while the extensive hyperparameter optimization sweeps for all models were executed on an NVIDIA DGX cluster equipped with A100 GPUs.

6.3.1 Human Performance Baselines

Human Evaluated SSC and Color Correlation

To establish an initial human performance baseline and assess the ability of experts to estimate ripeness from visual appearance, a specific validation test was performed. A set of 100 grape images was presented to a group of agronomists experienced in table grape ripeness estimation. For these specific images, a °Brix ground truth had been collected using the physical sampling technique described previously (Figure 6.4). The dataset consisted of standard JPG smartphone images of the Black Pizzutello variety collected during the 2021 harvest. The images were presented without additional context, and the experts were asked to assign a numerical value on a scale from 1 to 10 evaluating the degree of color `color_score`, the peduncle lignification stage `lignification`, and the berry shape `berry_shape`. Note that this 1 to 10 color scale differs from the 1 to 6 agronomic scale defined in Section 6.2.1. This experiment used a decimal scale strictly for the ease of human interpretation.

To create a binary harvesting prediction from this data, the agronomists' `color_score` was used as a proxy for their final decision by applying a threshold of 6. This thresholded prediction was designated as `harvesting_pred`. The correlation between these three visual estimates and the physically measured average °Brix `mean_brix` was then calculated to validate the core hypothesis. The resulting correlation matrix is shown in Figure 6.8.

The correlation score between the average °Brix and the expert color evaluation is 0.505, with $p < 0.001$ and a confidence interval of [0.342, 0.638]. The correlation between average °Brix and peduncle lignification is 0.337 ($p < 0.001$, $CI = [0.151, 0.501]$), while the correlation with berry shape is considerably lower at -0.044, which falls below any statistical significance threshold.

The observed significant correlation between the visual color assessment of human experts and actual measured °Brix values confirms that visual appearance is a valid proxy for ripeness. This suggests that a data-driven computer vision algorithm could theoretically achieve human-level accuracy in estimating grape quality for harvesting decisions.

This experiment also serves to establish a concrete human performance baseline. Given that a °Brix value greater than 17 is considered sufficient for harvesting,

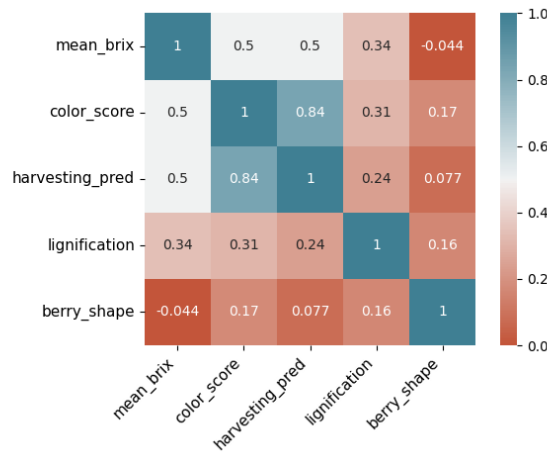


Figure 6.8. Correlation matrix of key parameters from the human evaluation study on 100 grape images. The matrix shows the relationship between the measured `mean_brix` and the scores assigned by experts for `color_score`, `lignification`, and `berry_shape`. The strong correlation of 0.505 between `mean_brix` and `color_score` validates the use of visual appearance for ripeness estimation.

the thresholded `mean_brix` was utilized as the ground truth harvesting label and compared against the experts' `harvesting_pred`. The human evaluators achieved a precision of 0.43 and a recall of 0.83. This translates to a false discovery rate (Type I error rate) of 0.57 and a false negative rate (Type II error rate) of 0.17. These values serve as the primary benchmark for the proposed algorithms.

In the context of automated harvesting, it is important to note that Type I errors (false positives) carry a significantly higher economic cost than Type II errors (false negatives). Ripe bunches left on the vine due to a false negative will eventually be harvested during a subsequent pass, whereas unripe bunches harvested due to a false positive must be discarded. The relatively high false discovery rate of the experts indicates that relying solely on a high color score results in a significant number of unripe bunches being collected due to natural mismatches between °Brix and skin color. Increasing the threshold for the `harvesting_pred` to 7 trades Type I errors for Type II errors, resulting in precision and recall scores of 0.35 and 0.37, respectively. This demonstrates that while °Brix and color are broadly correlated, the relationship is not strictly linear. The high number of borderline cases that human experts miss-classify highlights that training an appearance-based estimator is a non-trivial task.

Color Constancy effect on Human Evaluation of Color

As part of the exploratory analysis, the study investigated whether automatic camera post-processing (standard JPG) versus explicit color correction (white-balanced raw) significantly affects a human expert's visual evaluation. Agronomists were asked to label 50 pairs of images containing one standard and one white-balanced version of the same bunch, presented in a randomized sequence. A statistical analysis (z-test) of the assigned labels revealed no significant difference between the two groups

($p = 0.17$). This null finding confirmed that the two image types are perceptually similar to human experts, justifying the subsequent methodological approach of testing the computer vision models on both standard and white-balanced datasets to evaluate algorithmic robustness.

6.3.2 Histogram-based Model Evaluation

To establish a performance baseline for a computationally inexpensive approach, the histogram-based feature model was evaluated using Ridge Regression. These initial experiments were conducted on the **Simple** dataset splits, which contain 112 samples. The data was partitioned into a training set of 89 samples and a test set of 23 samples. The training images were augmented with horizontal flips, yielding a total of 178 training images. Both standard JPG and white-balanced images were evaluated to assess the influence of image post-processing on the handcrafted features.

Model performance and hyperparameters were assessed using a 5-fold cross-validation procedure on the training set. A grid search was conducted over key hyperparameters, including the number of horizontal and vertical histogram bins (n_{bin_x}, n_{bin_y}), the resulting feature vector length (**feature_len**), the choice of color space (RGB, HSV, or CIELAB), and the regularization strength (λ). The best-performing model achieved a validation Mean Absolute Error (MAE) of 1.46 °Brix. This result represents a notable 27% improvement over the naive statistical baseline for this dataset, which exhibited a Mean Absolute Deviation (MAD) of 1.99 °Brix.

Analysis of the hyperparameter search revealed several key insights into the model’s behavior. As illustrated in the feature correlation plot (Figure 6.9a), the regularization strength λ and the parameters defining the feature length exerted the most significant impact on performance. The analysis also confirmed the consistent benefit of utilizing the CIELAB color space, which generally outperformed standard RGB representations.

Furthermore, the effect of using standard JPG images versus color-corrected white-balanced (WB) images was explicitly investigated. Figure 6.9b demonstrates that models trained on standard JPG images consistently achieved a lower validation error than their counterparts trained on white-balanced images using identical hyperparameters. This suggests that for this specific feature-based approach, the camera’s default image signal processing pipeline preserves beneficial information, rendering the extra step of manual color correction unnecessary.

6.3.3 Multi-task DNN Model Evaluation

The multi-task DNN model was evaluated on the more challenging dataset splits to rigorously assess its performance under realistic covariate shifts. The **S21ToS22Phone** splits were utilized to test cross-seasonal robustness utilizing a consistent device type, while the **MixedSeason** splits (tested entirely on D435i images) were used to evaluate cross-device generalization, closely simulating a real-world deployment scenario.

The models were trained according to the splits described in Table 6.4. During training, in addition to the standard regression and classification metrics and the custom Validation BCH score used for cross-validation, the harvesting precision,

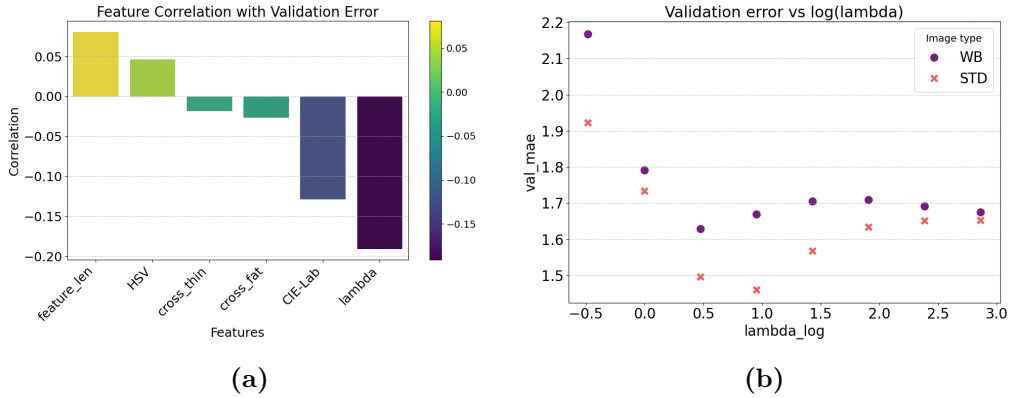


Figure 6.9. (a) Correlation between validation error (MAE) and hyperparameters for the top 10% performing runs, highlighting the direction of the correlation. Regularization strength (λ) and the use of the CIELAB color space exhibit the strongest correlation with improved performance (reduced error), whereas increased feature length correlates with higher validation error. (b) Test error for models with hyperparameters identical to the best model, but with varying λ values (logarithmic scale) and color correction. This isolates the effect of regularization and demonstrates that standard JPG images consistently yield better performance than color-corrected counterparts for this model.

recall, and F1-score were monitored using the same 17 °Brix threshold methodology applied to the human evaluators.

The multi-task model demonstrated strong performance and robust generalization capabilities. On the cross-season test set ($S21ToS22Phone_{STD}$), the best model achieved an MAE of 1.55 °Brix. Performance was even stronger on the highly challenging cross-device test set ($MixedSeason_{STD}$), achieving an MAE of 1.05 °Brix. These results significantly outperform their respective statistical baselines of 1.71 and 1.36 °Brix MAD (Table 6.4), highlighting the deep learning model’s ability to extract robust visual features for SSC estimation.

These top-performing models were identified through an extensive hyperparameter search guided by the custom BCH metric. A high correlation was observed between the minima of the BCH metric and the maxima of the F1-score for binary harvesting predictions. This indicates that while the two separate prediction branches could occasionally reach better individual performance during training, these isolated improvements did not reliably translate to better practical harvesting estimations. Because the BCH metric explicitly penalizes models that fail to meet minimum acceptable thresholds for either individual task, it effectively facilitated the selection of models where the joint learning task produced the most reliable overall predictions.

An important observation derived from the multi-task training process is that optimal performance for SSC estimation (regression) and color quality (classification) were rarely reached at the exact same training epoch. This highlights an inherent optimization trade-off when utilizing a single shared feature extractor for multiple correlated but distinct tasks. For specific agronomic applications, this implies that different model checkpoints could be selected depending on whether the primary operational goal is maximizing chemical SSC accuracy or visual color classification

accuracy.

Finally, the impact of color correction on the deep learning model’s performance was analyzed, mirroring the investigation conducted for the histogram features. To isolate this effect, the best-performing models on the standard `MixedSeasonSTD` dataset were compared against runs using identical hyperparameters but trained on the white-balanced `MixedSeasonWB` data. Figure 6.10 plots the results of this comparison against the log-ratio of the regularization weights (α/β). The resulting regression lines demonstrate that the performance difference between the two image distributions was minimal, with largely overlapping confidence intervals. This confirms that the proposed deep learning architecture is highly robust to inherent color variations and can effectively learn the relevant latent features without requiring complex, environment-specific color calibration steps prior to field deployment.

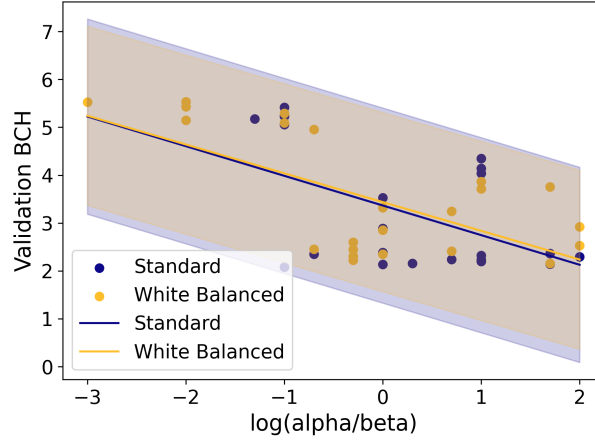


Figure 6.10. Linear regression analysis of runs utilizing hyperparameters similar to the top 20 models on the `MixedSeasonSTD` and `MixedSeasonWB` datasets. The shaded area represents the 95% confidence interval. While the models trained on color-corrected samples present slightly higher values on average, the smaller difference compared to Figure 6.9b suggests no practical advantage in using white-balanced raw images over standard JPG images in this context.

6.4 Discussion

The experimental results demonstrate that low-cost, in-field RGB imaging provides a viable alternative to complex and expensive hyperspectral systems for SSC estimation. By effectively capturing the latent correlations between external appearance and internal chemistry, the proposed methods overcome the physical instrumentation bottleneck that typically limits autonomous quality assessment. This section synthesizes the performance of these models, compares them against established human and statistical baselines, and discusses the practical implications for robotic deployment.

6.4.1 Summary of Performance and Comparison

Table 6.5 consolidates the performance of the proposed models against the baselines and other non-contact methods documented in the literature. The multi-task deep learning model successfully learned meaningful representations from the visual data, achieving a best MAE of 1.55 °Brix on the cross-season **S21ToS22Phone** dataset and an even lower 1.05 °Brix on the highly challenging cross-device **MixedSeason** dataset. These results significantly outperform the naive statistical baselines of 1.71 and 1.36 °Brix MAD, respectively.

Table 6.5. Master comparison of model performance against baselines and state-of-the-art methods. Note that literature results are derived from different grape varieties and environmental conditions, and RMSE results are presented for contextual comparison.

Method	Dataset / Conditions	Performance Metric	Result
Statistical Baseline	S21ToS22Phone Test	MAD	1.71 °Brix
Statistical Baseline	MixedSeason Test	MAD	1.36 °Brix
Human Experts	100 JPG images from the 2021 harvest	Precision / Recall	0.43/0.83
Our Histogram Model	Simple _{STD}	MAE	1.46 °Brix
Our DNN Model	S21ToS22Phone _{STD}	MAE	1.55 °Brix
Our DNN Model	MixedSeason _{WB}	MAE	1.05 °Brix
Our DNN Model	Harvesting Task on S21ToS22Phone	Precision / Recall	0.91/0.84
Our DNN Model	Harvesting Task on MixedSeason	Precision / Recall	0.62/0.83
[49] HSI	Tempranillo / In-Field	RMSE	1.27 °Brix
[12] RGB + NN	Tempranillo / In-Field	RMSE	4.63 °Brix
[158] RGB + MLR	Kyoho / Laboratory	RMSE	0.78 °Brix

To translate these continuous regression errors into actionable agronomic metrics, the DNN model was evaluated on the binary harvesting task. As shown in Table 6.5, the model achieved a precision and recall of 0.91 and 0.84 on the cross-season split, and 0.62 and 0.83 on the cross-device split. This compares favorably to the human expert baseline of 0.43 precision and 0.83 recall. Maintaining a high recall while simultaneously achieving higher precision is particularly valuable in a real-world harvesting scenario because it directly reduces the costly error of collecting unripe grapes. Furthermore, the experiments consistently showed that the F1-score computed from these precision and recall values is highly correlated with the custom BCH metric. This suggests that the model’s high performance derives from the effective, joint learning of features related to both color classification and SSC regression.

When benchmarked against alternative visual estimation techniques, this low-cost RGB approach demonstrates highly competitive performance. Direct comparisons are inherently challenging due to differing grape varieties, environmental conditions, and the specific error metrics reported in the literature. Although RMSE mathematically upper-bounds MAE, meaning direct numerical comparisons are not perfectly equivalent, the best achieved MAE of 1.05 °Brix compares very favorably against reported RMSE values from in-field HSI systems, which typically range from 1.27 to

2.20 °Brix. Furthermore, it represents a substantial improvement over other in-field RGB methods that report errors closer to 4.63 °Brix RMSE, effectively approaching the accuracy of methods performed under strictly controlled laboratory conditions.

6.4.2 Practical Implications and Trade-offs

This chapter presents two distinct solutions tailored for different operational constraints within an agricultural robotics context. The histogram-based model, while less accurate overall, offers significant advantages in computational efficiency and simplicity. Once the model is trained, the estimation process reduces to a simple dot product between the extracted feature vector and the estimator weights. The most computationally expensive operation in this pipeline is the pixel-wise integral sum required for the histogram computation, which scales linearly with the number of pixels. However, as demonstrated in the experiments, the low spatial frequency of the relevant visual features allows this method to operate effectively on very low-resolution images, drastically reducing the computational burden. Consequently, the low memory and processing footprint of the histogram features make this approach highly suitable for real-time deployment on resource-constrained edge devices or as a rapid, first-pass filter on a robotic platform.

In contrast, the multi-task DNN model provides superior accuracy and robustness to covariate shifts, particularly across different camera devices and growing seasons, but at a greater computational cost. This makes the deep learning approach ideal for applications where peak performance is paramount, such as a centralized quality control system or deployment on advanced robotic platforms equipped with dedicated GPU hardware.

6.4.3 Limitations and Future Work

Several limitations exist that provide avenues for future research. This study was conducted exclusively on a single black grape variety (Pizzutello Nero). While focusing on a single variety is common in the literature, extending these methods to white grape varieties requires specific adjustments. In black or red grape varieties, the ripening color gradient is primarily perceived as a contrast and intensity change. For white varieties, ripening manifests as subtle hue and saturation changes. Detecting these specific variations reliably across shifting lighting conditions will likely require a tightly color-calibrated RGB imaging system or the persistent inclusion of a calibration pattern within the camera’s field of view. Additionally, while the proposed methods demonstrated robustness to the uncontrolled, natural lighting of the field across different seasons, further experiments under more extreme weather conditions (such as direct, bright sunlight versus heavy cloud cover) are needed to better quantify the operational domain of the system. Future work will focus on expanding the dataset to include these variations and exploring multi-sensor fusion techniques to further improve robustness.

6.5 Concluding Remarks

This chapter successfully demonstrated the feasibility of estimating Soluble Solid Content (SSC) in table grapes using low-cost RGB sensors under challenging, uncontrolled field conditions. By expanding the perception pipeline beyond external surface geometry to infer internal chemical properties, this work addresses a critical requirement for fully autonomous robotic harvesting.

To overcome the physical annotation bottleneck associated with acquiring chemical ground truth, a novel multi-season, multi-device dataset of Black Pizzutello was collected and annotated for both SSC and agronomic color quality. This dataset established a framework for validating data-driven models against the covariate shifts inherent to agricultural environments, specifically seasonal variations and differing imaging hardware.

Two algorithmic approaches were proposed and validated. The first, a lightweight method utilizing classical handcrafted histogram features, proved highly efficient for devices with strictly limited computational power. The second, a multi-task deep neural network, effectively leveraged the latent correlations between visual appearance and SSC. Driven by a custom Brix-Color-Hinge (BCH) metric designed to enforce agronomic quality standards, this deep learning architecture achieved reliable SSC estimation with a significant performance margin compared to average human evaluators. Achieving a best-performing MAE of 1.05 °Brix on a challenging cross-device test set, the model proved highly competitive with expensive in-field hyperspectral systems and successfully surpassed human harvesting decisions with high precision and recall. Ultimately, the demonstrated robustness to seasonal and hardware-induced covariate shifts positions these RGB-based solutions as viable, cost-effective perception modules for agricultural robotics.

This chapter also concludes the investigation into precision agriculture. Throughout these applications, from robust instance segmentation to synthetic anomaly generation and chemical ripeness estimation, a persistent challenge has been the annotation bottleneck. The methodologies developed thus far demonstrate that this limitation can be effectively mitigated using surrogate supervision and compositional synthesis, reducing the reliance on massive, manually curated datasets in proximal, object-centric environments.

However, data inefficiency is not limited to localized field robotics. Macroscopic, scene-centric domains, such as Earth Observation, present parallel challenges compounded by vast geographic diversity and severe class imbalances. The second part of this dissertation addresses this broader context. The following chapters investigate how the core principles of compositional synthesis and semi-supervised learning can be generalized and transferred to satellite imagery to address complex tasks such as land use land cover classification and change detection.

Part II

Remote Sensing

Chapter 7

Cut-and-Paste Augmentation for Satellite Semantic Segmentation

7.1 Introduction

The deployment of deep learning models for Earth Observation is fundamentally constrained by the annotation bottleneck. Semantic segmentation in satellite imagery involves classifying each pixel into categories like impervious surfaces, cultivated areas, forests, or water bodies. This process, often referred to as Land Use Land Cover (LULC) classification, is a foundational task for interpreting the Earth’s surface and for applications such as environmental monitoring and urban planning. However, developing robust models for this task is limited by the availability of labeled data [135], the inherent variability and complexity of satellite imagery, and pronounced class imbalances within existing datasets.

Generating pixel-level annotations across vast geographic areas is costly, labor-intensive, and prone to distributional biases. Consequently, minority classes are consistently underrepresented, leading to models that struggle with rare but significant geographic features.

To mitigate these issues, traditional approaches have relied on image-level augmentations, generating new training samples through photometric and geometric transformations such as flips, rotations, crops, and color adjustments to improve model generalization. Beyond these standard image-level transformations, instance-level augmentations, such as those based on the Cut-and-Paste concept [38, 122, 39, 44], have shown promise for object detection and instance segmentation tasks in conventional camera imagery, as well as in construction resource detection using UAV-acquired images [8]. These methods embody the principle of compositional synthesis: by manipulating instance placement, they introduce diverse object representations into the training distribution [38]. While some techniques refine this concept by placing object masks in realistic locations to create consistent appearances [122, 36, 39, 37], others paste instances randomly without strict contextual modeling [44].

In the specific context of satellite imagery, generative adversarial networks have been employed to paste buildings into high-resolution optical images to aid binary change detection models [19]. Another approach involves directly pasting buildings

onto diverse backgrounds to enhance building segmentation in very high-resolution images [60]. Despite these advancements, a gap remains in adapting these compositional techniques to the specific challenges of semantic segmentation in satellite imagery. The primary obstacle is that standard semantic labels do not differentiate between individual objects or instances, which are a prerequisite for instance-level augmentations. Moreover, the application of Cut-and-Paste augmentation techniques in medium-resolution satellite images (1 to 10 meters per pixel) remains largely unexplored.

To address these limitations, this chapter extends the compositional synthesis strategy to the satellite domain by exploring a Cut-and-Paste augmentation strategy for semantic segmentation of satellite images. The approach addresses the lack of instance-level annotations by extracting and saving discrete instances from the semantic segmentation labels using connected components analysis. This step enables the generation of augmented images by pasting the extracted instances in various configurations onto the original training images. Unlike previous approaches that restrict the source of instances to the target image itself [122, 39] or to another randomly selected image from the dataset [19], the proposed method leverages the entire training set to allow for more diverse compositions. Overlapping instances are also permitted, creating atypical yet plausible scenarios that encourage the models to generalize beyond frequent spatial patterns.

This technique can be categorized as model-free, multi-image, and instance-level as defined in [161] because it does not rely on additional generative models, utilizes multiple images to produce augmented samples, and focuses on manipulating specific objects. This method provides a viable solution for generating new semantic segmentation data without requiring additional manual annotations, thereby reshaping the training distribution to improve robustness.

The experimental evaluation, conducted on the DynamicEarthNet dataset using a standard U-Net baseline model, demonstrates that the proposed Cut-and-Paste augmentation improves the mean Intersection over Union (mIoU) score on the test set from 37.9 to 44.1, confirming its utility in mitigating the annotation bottleneck in remote sensing.

7.2 Methodology

The proposed augmentation pipeline consists of two primary stages, as illustrated in Figure 7.1. First, an offline preprocessing step is conducted to extract the instances of interest from the training set. Second, an online augmentation phase randomly pastes these extracted objects onto images during the training loop.

7.2.1 Instance Extraction

The initial stage adapts the Cut-and-Paste augmentation paradigm for semantic segmentation. Unlike instance segmentation, where labels explicitly delineate individual objects, semantic segmentation datasets, such as DynamicEarthNet [140], provide class information exclusively at the pixel level, merged into a single composite mask. To leverage Cut-and-Paste techniques in this context, the composite mask must

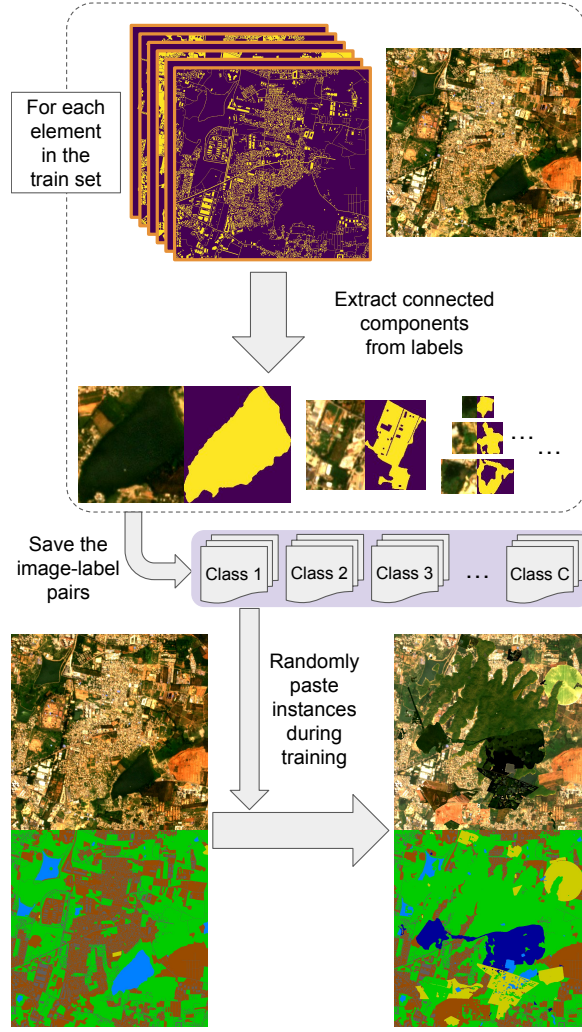


Figure 7.1. Overview of the proposed Cut-and-Paste augmentation technique. In the initial offline phase, connected components analysis extracts discrete class instances from the semantic labels to populate an instance bank. During the online training phase, these instances are randomly sampled and pasted onto the background images.

first be separated into distinct instance masks. This step enables the independent manipulation and pasting of instances in the subsequent stage.

To identify individual objects within an image, the methodology relies on connected components analysis, grouping contiguous pixels that belong to the same semantic class. For each image I , its corresponding label L is separated into C binary masks L_c , one for each class. The connected components algorithm is then applied to each binary mask L_c to extract its respective N_c discrete components. In practice, this corresponds to partitioning the composite label into a set of non-overlapping mask instances $L_{c,i}$, each belonging to one of the C classes:

$$L = \bigcup_{c=1}^C L_c = \bigcup_{c=1}^C \bigcup_{i=1}^{N_c} L_{c,i} \quad (7.1)$$

The resulting extracted objects consist of cropped multispectral image patches and their corresponding binary masks $(I_{c,i}, L_{c,i})$. These pairs are stored to construct a comprehensive bank of instances for each semantic category.

7.2.2 Instance Pasting

The second stage operates dynamically during the training process, augmenting each image-label pair on the fly using the instance bank generated in the previous step. For a given training image, a semantic category c is uniformly sampled, and an instance $(I_{c,i}, L_{c,i})$ belonging to that class is randomly selected. This chosen instance undergoes optional geometric augmentations, such as flips and rotations, before being pasted onto the current training sample at a random location.

This pasting process is repeated N times per training image, where N acts as a tunable hyperparameter defining the volume of injected objects. In contrast to existing approaches [19], this method permits overlapping instances, which can occasionally generate unusual spatial configurations. Although such configurations might be statistically rare, they represent plausible real-world physical overlaps and encourage the trained model to generalize beyond common structural patterns.

Once the pasting iteration is complete, the final composite image undergoes standard non-destructive transformations, specifically horizontal and vertical flips, as well as 90-degree rotations, to further expand the augmentation space.

7.3 Experiments

To validate the efficacy of the proposed Cut-and-Paste augmentation in mitigating the annotation bottleneck, a comprehensive set of experiments was conducted. This section describes the dataset, the evaluation metrics, the training protocol, and an analysis of the results.

7.3.1 Dataset

The experimental evaluation utilized the DynamicEarthNet dataset [140], which consists of daily, multispectral satellite imagery acquired by PlanetLabs. This dataset encompasses 75 diverse areas of interest globally, with temporal coverage spanning 24 months from January 2018 to December 2019. Each image features both RGB and near-infrared bands at a spatial resolution of 3 meters per pixel. The dataset provides monthly pixel-wise semantic segmentation labels across seven classes: impervious surface, agriculture, forest & other vegetation, wetlands, soil, water, and snow & ice. Following the established methodology outlined in [140], the snow & ice class was excluded from this study due to its severe underrepresentation. Since the focus is on fully supervised semantic segmentation, only images possessing a corresponding ground truth label (the first day of each month) were utilized.

The dataset is natively partitioned into training, validation, and test sets. However, because the official validation and test labels are withheld and require submission to a closed evaluation server¹, an internal validation strategy was necessary for

¹Codalab platform: <https://codalab.lisn.upsaclay.fr/competitions/2882>

hyperparameter tuning. The official training set was subdivided into 1200 images for model training and 120 images for internal validation. To strictly prevent spatial data leakage, this split was carefully stratified to ensure that any specific geographic area of interest appeared in only one of the two subsets while maintaining a representative distribution of all semantic classes in both.

7.3.2 Evaluation Metric

To assess the performance of the segmentation models, the mean Intersection over Union (mIoU) was employed as the primary quantitative metric. This metric computes the average of the intersection over union values between the predicted masks and the ground truth labels across all considered classes:

$$mIoU = \frac{1}{C} \sum_{i=1}^C \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \quad (7.2)$$

where C represents the total number of classes, A_i denotes the ground truth mask for class i , and B_i represents the predicted mask for class i .

7.3.3 Training Details

Since the objective is to isolate the specific impact of the augmentation strategy rather than to achieve state-of-the-art performance through architectural complexity, the experiments employ a standard U-Net model [124]. The U-Net architecture was trained from scratch using the Adam optimizer with a learning rate of 1e-4 for 200 epochs. Model weights were saved at the epoch achieving the highest mIoU on the internal validation set. To ensure statistical reliability and reproducibility, all hyperparameters remained fixed after the initial baseline optimization, and each experimental configuration was executed three times. The reported metrics reflect the mean and standard deviation across these runs.

The baseline model was trained using standard geometric augmentations, specifically horizontal flips, vertical flips, and random 90-degree rotations. The Cut-and-Paste augmentation was subsequently introduced in addition to these standard transformations to explicitly measure its contribution. The impact of the proposed augmentation was evaluated by varying the number of pasted instances per image ($N \in \{10, 100, 1000\}$). Additionally, the experiments explored the influence of applying geometric transformations (flips and rotations) to the individual instances prior to pasting them onto the image. These are referred to as pre-pasting augmentations.

7.3.4 Results

Table 7.1 presents the performance of the models on both the official validation set and internal validation set. The results demonstrate that the application of Cut-and-Paste augmentation yields notable improvements over the baseline.

When the augmentation was applied conservatively with ten instances ($N = 10$), the mIoU on the official validation set improved, accompanied by a negligible decrease on the internal validation set. Increasing the injection rate to 100 instances ($N = 100$) resulted in substantial performance gains across both validation splits,

indicating that the method scales effectively by providing a denser supervisory signal for the minority features. However, when the volume was increased to 1000 instances ($N = 1000$), the mIoU gains diminished. This performance drop suggests an upper limit to the augmentation capacity, beyond which extreme object crowding obscures the natural spatial context of the underlying image and degrades the learning process.

Table 7.1. Validation performance of the Cut-and-Paste augmentation compared to the baseline U-Net model. The evaluation considers the effect of varying the number of injected instances (N) and the inclusion of pre-pasting geometric transformations. Metrics represent the mean and standard deviation of the mIoU computed over three independent training runs.

	Pre-pasting augmentations	Val mIoU	Internal Val mIoU
Baseline		34.1 (2.1)	43.5 (1.0)
C&P N=10	✓	34.5 (0.8)	43.4 (0.9)
C&P N=100	✓	35.9 (1.1)	47.4 (0.7)
C&P N=1000	✓	35.6 (0.6)	47.0 (0.6)
C&P N=10	✗	35.5 (1.3)	43.2 (0.9)
C&P N=100	✗	36.0 (1.5)	47.5 (0.8)
C&P N=1000	✗	34.8 (2.0)	47.0 (0.7)

The most effective overall configuration utilized 100 pasted instances without pre-pasting augmentations, achieving the highest average validation mIoU of 36.0. The configuration using the same number of instances but incorporating pre-pasting augmentations performed competitively, reaching a validation mIoU of 35.9. This finding indicates that pre-pasting augmentations are not strictly necessary for achieving substantial performance gains, as the model can benefit purely from the structural diversity introduced by the raw Cut-and-Paste process itself.

Figure 7.2 provides a qualitative visual comparison of the segmentation results between the baseline and the best-performing model on sample images from the internal validation set. The visual improvements are particularly evident within the impervious surface and agriculture classes, where the augmented model demonstrates enhanced boundary delineation and predictive accuracy.

For the final evaluation on the withheld test set (Table 7.2), only the two best configurations were assessed. In this evaluation, the performance gains were even more pronounced. The configuration pasting 100 instances without pre-pasting augmentations achieved a test mIoU of 44.1, while the variant with pre-pasting augmentations reached 42.3. Both configurations significantly surpassed the baseline mIoU of 37.9. These results confirm the fundamental efficacy of the Cut-and-Paste augmentation technique for remote sensing. They also highlight that hyperparameters governing the injection density and instance-level transformations must be carefully calibrated to prevent scene degradation and maximize the learning signal.

7.4 Concluding Remarks

This chapter addressed the annotation bottleneck in the Earth Observation domain by extending the compositional synthesis paradigm to satellite imagery. In particular, it

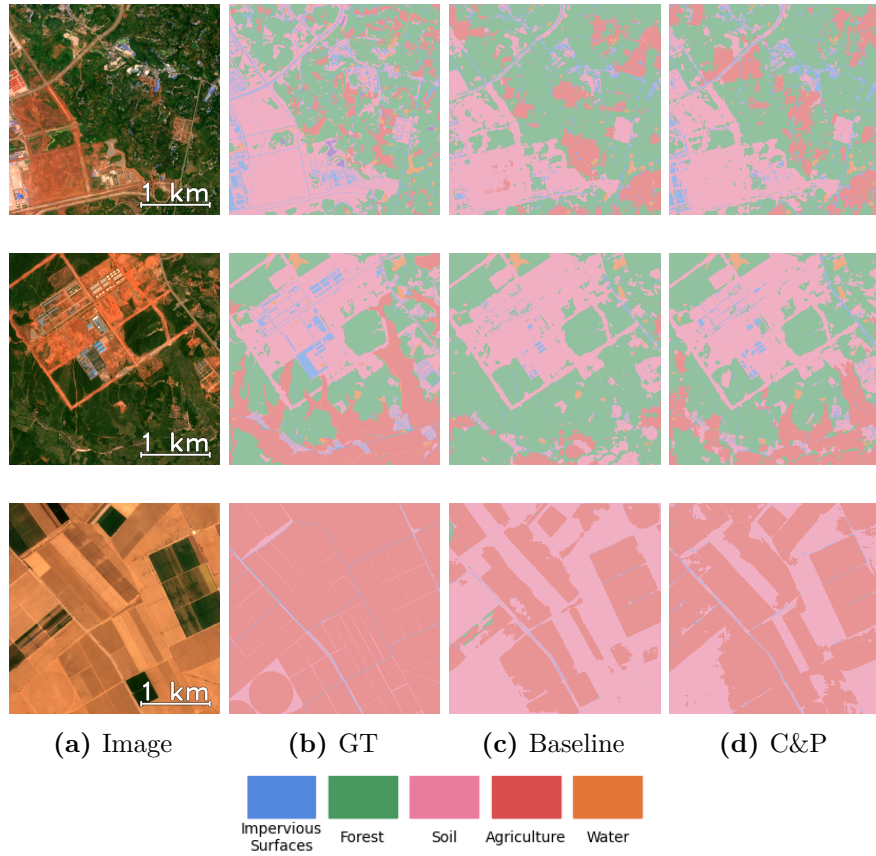


Figure 7.2. Qualitative visualization of semantic segmentation results on the internal validation set. (a) Original multispectral satellite imagery, converted to RGB and enhanced for visibility. (b) Corresponding ground truth segmentation labels. (c) Predictions generated by the baseline U-Net model. (d) Predictions generated by the U-Net model utilizing the optimal Cut-and-Paste augmentation configuration ($N = 100$, without pre-pasting augmentations).

explored a Cut-and-Paste augmentation strategy adapted for semantic segmentation.

Because standard semantic labels lack the instance-level granularity typically required for object-level transformations, the proposed methodology leveraged connected components analysis to automatically isolate and extract discrete objects from the composite masks. By dynamically pasting these instances into various configurations during the training loop, the approach successfully generated highly diverse training samples without incurring the cost of additional manual annotation. This process effectively reshaped the training distribution, explicitly increasing the representation of minority classes and encouraging the model to learn robust, generalizable features rather than overfitting to frequent spatial layouts.

The experimental validation demonstrated the practical value of this approach. Implementing the augmentation on a standard U-Net baseline model yielded a substantial performance increase on the DynamicEarthNet test set, raising the mean Intersection over Union from 37.9 to 44.1. This confirms that intelligently expanding the training distribution through compositional synthesis is an effective strategy for

Table 7.2. Test set performance evaluating the optimal Cut-and-Paste augmentation configurations ($N = 100$) against the baseline model.

	Pre-pasting augmentations	Test mIoU
Baseline		37.9
C&P $N=100$	✓	42.3
C&P $N=100$	✗	44.1

mitigating data scarcity in remote sensing.

However, while semantic segmentation addresses the static state of the Earth’s surface, many high-impact applications require monitoring how geographic features evolve over time. The fundamental challenge of class imbalance is significantly more severe in the context of Change Detection, where actual change events often constitute a minute fraction of the observed pixels. Consequently, the following chapter investigates how targeted data synthesis and semi-supervised learning can address these inherent data constraints.

Chapter 8

ChangeMix: Change-Aware Cut-and-Paste Augmentation

8.1 Introduction

Monitoring the dynamic evolution of the Earth’s surface introduces significant data challenges, primarily the inherent scarcity of change events. This rarity presents a primary obstacle for Change Detection (CD), the task of identifying transitions in the state of an area observed at different times, which supports various applications, including monitoring urban expansion, managing land use, assessing the impact of natural disasters, updating geographic databases, and tracking environmental changes [150, 64].

While traditional methods relied on hand-crafted features, the field has benefited from advances in deep learning [130]. Modern architectures, particularly those based on Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) [106], demonstrate a strong capacity for learning hierarchical, discriminative features directly from data. However, the efficacy of these data-driven models depends on the availability of large-scale, high-quality labeled datasets. Creating pixel-level annotations for bi-temporal images is labor-intensive and requires expert knowledge to distinguish meaningful semantic changes from irrelevant variations caused by seasonal shifts or sensor noise [150]. This annotation bottleneck often leads to overfitting and reduced generalization when labeled data is scarce.

To address these limitations, Semi-Supervised Learning (SSL) has emerged as a standard paradigm, enabling models to learn from limited labeled data alongside abundant unlabeled data. The current landscape of Semi-Supervised Change Detection (SSCD) primarily adapts consistency regularization and pseudo-labeling from general computer vision. Consistency regularization enforces the stability of model predictions under various input perturbations, while pseudo-labeling generates provisional labels from unlabeled data to augment training. Frameworks such as FixMatch [134] and UniMatch v2 [163] effectively combine these mechanisms for semantic segmentation.

Despite their success in semantic segmentation, these general SSL frameworks often struggle when directly applied to CD due to two domain-specific challenges that they do not explicitly address: extreme class imbalance and temporal bias.

Unlike general scenes where objects may be more evenly distributed, CD datasets typically exhibit a long-tailed distribution where "unchanged" pixels vastly outnumber "changed" ones. Frequently, more than 95% of the pixels in a given dataset represent no change [114, 57, 83]. Standard mixing augmentation strategies, such as CutMix, randomly sample regions from the co-registered bi-temporal image pair that are statistically likely to contain only background pixels, as illustrated in Figure 8.1. Consequently, the model becomes biased toward predicting "no change", often confusing true change events with noise. Moreover, real-world datasets frequently show an asymmetry between change types, with object appearance (such as construction) occurring more often than disappearance (such as demolition) [131]. This asymmetry introduces a temporal bias, where model performance varies depending on the temporal ordering of the image pair. Standard augmentations do not account for the temporal directionality of bi-temporal pairs, which can exacerbate this bias. As a result, existing SSCD methods lack a mechanism to explicitly inject targeted supervision for the minority "changed" class, limiting their practical effectiveness.

To address these issues, this chapter proposes *ChangeMix*, a change-aware augmentation designed to extend the compositional synthesis paradigm to semi-supervised change detection. Unlike standard region-mixing methods like CutMix [164] that sample regions randomly, ChangeMix builds upon the Cut-and-Paste concept explored in Chapter 7 to serve as a targeted sampling strategy within the SSL pipeline. It features a dynamic Instance Bank mechanism that collects "changed" instances during training as paired pre- and post-change crops, together with their corresponding mask and confidence scores. These instances are accumulated from both the limited ground truth and the confident pseudo-labels derived from unlabeled data. The augmentation then retrieves instances from the bank and pastes them onto strongly augmented unlabeled images. This process injects targeted supervision for the minority class directly into the student model's training loop, augmenting the learning signal to mitigate class imbalance. To address temporal bias, ChangeMix randomly swaps the temporal order of pre- and post-change patches before pasting, simulating appearance and disappearance events with equal probability.

ChangeMix is evaluated across multiple label budgets (5% to 40%) on the LEVIR-CD [20] and WHU-CD [63] benchmarks. The results demonstrate that the proposed strategy is both framework-agnostic and model-agnostic, yielding consistent improvements over baselines using the FixMatch [134], UniMatch V1 [162], and UniMatch V2 [163] frameworks, as well as across DPT with DINOv2 [120, 106] (Small and Base) and DeepLabV3+ [21] architectures.

The main contributions presented in this chapter are summarized as follows:

1. The introduction of *ChangeMix*, a change-aware augmentation for SSCD that mitigates class imbalance by injecting targeted supervision for the minority "changed" class.
2. The formulation of a dynamic Instance Bank that accumulates change instances from both labeled data and filtered pseudo-labels, allowing the augmentation to evolve alongside the model during training.
3. The implementation of a temporal swapping strategy within the augmentation pipeline to mitigate temporal bias, balancing the representation of appearance

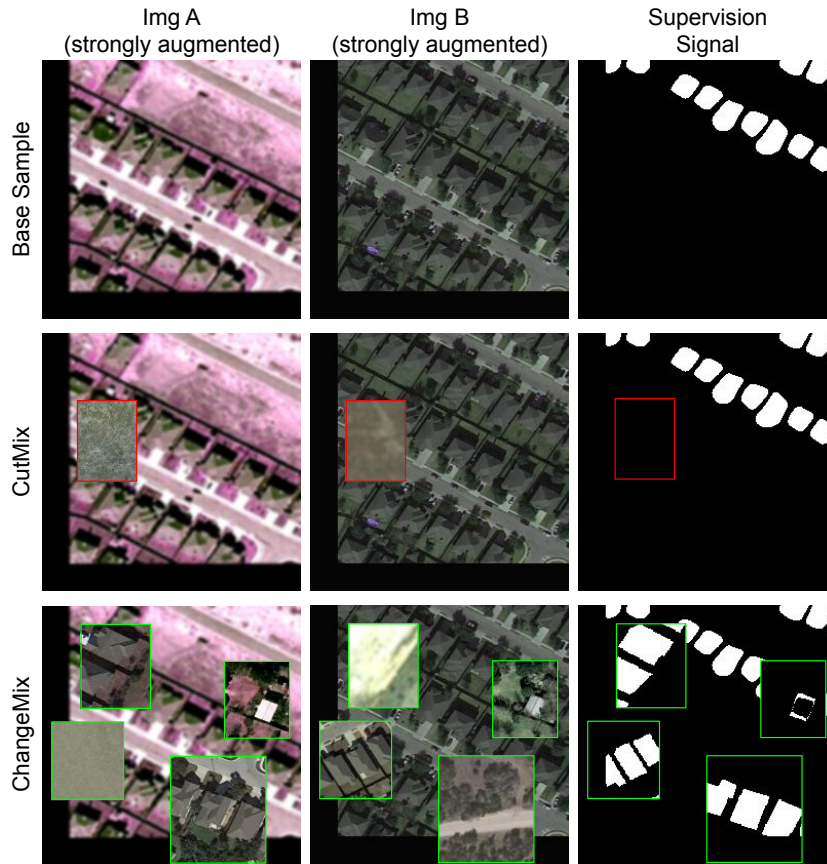


Figure 8.1. Comparison of augmentation strategies and their impact on supervision. The columns display the strongly augmented bi-temporal inputs of the same site (Img A and Img B) alongside the corresponding target supervision signal (pseudo-label). Top row (Base Sample): A typical training sample consists primarily of background pixels, offering sparse supervision for the "changed" class. Middle row (CutMix): Standard CutMix (red box) randomly samples a region. Due to the relative sparsity of changes, it statistically selects background content, resulting in empty supervision that reinforces the class imbalance. Bottom row (ChangeMix): The proposed method (green boxes) injects multiple high-confidence change instances from an Instance Bank. This densifies the positive supervision signal and introduces balanced appearance and disappearance events to mitigate temporal bias.

and disappearance events.

4. An evaluation of ChangeMix demonstrating its adaptability, showing consistent improvements over SSCD baselines across multiple architectures and SSL frameworks.

8.2 Related Work

8.2.1 Deep Learning-based Change Detection

Early change detection methods primarily relied on hand-crafted features and algebraic operations such as image differencing and Change Vector Analysis (CVA) [64]. The integration of deep learning has significantly advanced the field, enabling the end-to-end learning of hierarchical, discriminative features [130]. Initial approaches employed CNNs in Siamese architectures to extract and fuse bi-temporal features [31, 64]. Subsequent advancements introduced more sophisticated architectures, such as UNet++ [115], HRNet [148], and SNet [40], to preserve multi-scale spatial details.

More recently, Transformer-based architectures have demonstrated notable performance by modeling long-range dependencies. For instance, ChangeFormer [7] utilizes a hierarchical Transformer encoder to capture multi-scale temporal changes, while architectures like SwinSUNet [168] leverage Swin Transformer blocks for global feature extraction. Despite these architectural advances, the performance of deep learning models remains heavily dependent on large volumes of pixel-wise annotations, which are scarce and costly to acquire. This dependency has spurred increasing interest in learning with limited supervision.

8.2.2 Semi-Supervised Change Detection

SSL mitigates data scarcity by leveraging unlabeled data alongside a small labeled set. Current SSL methods predominantly rely on two paradigms: self-training and consistency learning. Self-training approaches generate pseudo-labels for unlabeled data to guide the training process, often employing thresholding to select reliable labels [149, 59]. Consistency-based methods enforce stability in model predictions under various perturbations. FixMatch [134] established a weak-to-strong paradigm where predictions derived from weakly augmented inputs serve as pseudo-labels for strongly augmented views. UniMatch [162] introduced dual strong perturbation streams to broaden the augmentation space, a concept later refined in UniMatch V2 [163], to accommodate ViT architectures. Concurrently, the integration of Visual-Language Models (VLMs) has emerged as a strategy for injecting high-level semantic guidance as additional supervision for unlabeled data. [58].

While these frameworks have achieved success in semi-supervised semantic segmentation, the remote sensing community has actively explored adapting these paradigms to the CD task. Early efforts often used adversarial learning frameworks to align the feature distributions of labeled and unlabeled data [113]. More recently, the field has converged toward consistency regularization. Bandara and Patel [6] proposed enforcing consistency under small random perturbations applied to the deep feature difference map. Similarly, WACS-SemiCD [114] emphasized that diverse perturbations are fundamental for exploiting unlabeled data, introducing a framework that integrates feature and sample-level consistency with class-balanced sampling. UniMatch [162] and UniMatch V2 [163] also investigated the application of their respective frameworks to the CD domain. Alternatively, SemiCD-VL [85] explored integrating VLMs to guide the learning process with high-level semantic abstractions.

8.2.3 Data Augmentation for Change Detection

Data augmentation is a fundamental component of modern SSL pipelines, enhancing model robustness and generalization. Beyond standard geometric and photometric transformations, region-based mixing strategies such as CutMix [164], Copy-Paste [44], and MixUp [169] have proven effective in general vision tasks and have recently been adapted for remote sensing applications [53, 131, 101].

However, standard mixing augmentations sample regions randomly, ignoring the underlying semantic content [169, 164]. In CD tasks, where the target objects are sparse, random sampling predominantly mixes background with background. This fails to provide sufficient supervision for the minority class and does not address the temporal bias favoring appearance over disappearance.

To mitigate these issues, recent works have focused on targeted mixing strategies. WACS-SemiCD [114] utilizes ground truth masks to paste labeled objects into unlabeled images. While this strategy is useful for domain adaptation, it is constrained by the limited diversity of the labeled set and cannot exploit the wider semantic variety present in the unlabeled data. Alternatively, CutMix-CD [131] improves upon generic mixing by using difference-based saliency maps to spatially target changed regions within unlabeled data. However, this approach is restricted to the information available within the current training batch, which limits augmentation diversity in batches where change instances are sparse or absent. Furthermore, by targeting a single salient region per image based on peak activation, it restricts the ability to significantly increase the sampling frequency of the minority class.

The proposed ChangeMix augmentation addresses these limitations by introducing a confidence-filtered Instance Bank and a multi-instance pasting strategy. The instance bank accumulates a diverse pool of change examples from both labeled ground truth and high-confidence pseudo-labels, ensuring a rich supervisory signal even when the current batch lacks sufficient change events. By pasting multiple non-overlapping instances per image and explicitly swapping temporal orders, the method densifies the supervision for the minority class and counteracts temporal bias.

8.3 Methodology

This section details the proposed ChangeMix augmentation and its integration into SSL pipelines for bi-temporal remote sensing CD. The problem definition is formalized first, followed by the description of a generalized weak-to-strong consistency framework that encompasses methods like FixMatch, UniMatch V1, and UniMatch V2. Finally, a comprehensive description of the ChangeMix augmentation strategy is provided.

8.3.1 Problem Definition

Given a pair of co-registered images $(I^A, I^B) \in \mathbb{R}^{C \times H \times W}$ acquired at times t_A and t_B respectively (where $t_A < t_B$), the task is to predict a pixel-wise change map $M \in \{0, 1\}^{H \times W}$ indicating unchanged (0) versus changed (1) categories. The

formulation assumes the availability of a small labeled set:

$$\mathcal{D}_L = \{(I_i^A, I_i^B, M_i)\}_{i=1}^{N_L}, \quad (8.1)$$

and a much larger unlabeled set:

$$\mathcal{D}_U = \{(I_j^A, I_j^B)\}_{j=1}^{N_U}, \quad (8.2)$$

where $N_U \gg N_L$. The objective is to train a CD model f_θ that can accurately predict change maps by leveraging both labeled and unlabeled data.

8.3.2 Generalized Weak-to-Strong SSL Framework

To demonstrate the versatility of the proposed approach, ChangeMix is formulated within a generic weak-to-strong consistency framework. This paradigm serves as the backbone for state-of-the-art methods such as FixMatch [134], UniMatch V1 [162], and UniMatch V2 [163]. The framework utilizes a student model f_θ optimized via gradient descent and a teacher model $f_{\bar{\theta}}$ used to generate pseudo-labels.

Teacher Model Update Strategy

The relationship between the student parameters θ and teacher parameters $\bar{\theta}$ depends on the specific SSL framework:

- **EMA Teacher (e.g., UniMatch V2, FixMatch):** The teacher model is an exponential moving average (EMA) of the student model to provide stable predictions:

$$\bar{\theta}_{t+1} = \alpha_t \bar{\theta}_t + (1 - \alpha_t) \theta_t, \quad (8.3)$$

where the EMA decay rate grows from 0 at iteration $t = 0$ towards an upper bound $\alpha_{max} = 0.996$ according to:

$$\alpha_t = \min \left(1 - \frac{1}{t+1}, \alpha_{max} \right) \quad (t \geq 0). \quad (8.4)$$

This scheduling strategy ensures that the teacher model gradually stabilizes while maintaining sensitivity to improvements in the student model.

- **Shared Weights (e.g., UniMatch V1):** The teacher model shares weights entirely with the student model ($\bar{\theta} = \theta$).

ChangeMix is compatible with both strategies, as it relies only on the availability of high-quality pseudo-labels $\hat{M}$ from the teacher model.

Consistency Regularization with ChangeMix

The primary component of the framework is the consistency loss enforced on unlabeled data. For an unlabeled pair (I^A, I^B) , the pipeline generates two types of views:

- A **weakly augmented view** (I_w^A, I_w^B) . This view includes only geometric transformations (resize, crop, and flip) and is processed by the teacher model $f_{\hat{\theta}}$ to generate pseudo-labels $\hat{M}$ and confidence scores $\hat{C}$:

$$\hat{M} = \arg \max f_{\hat{\theta}}(I_w^A, I_w^B), \quad (8.5)$$

$$\hat{C} = \max \text{softmax}(f_{\hat{\theta}}(I_w^A, I_w^B)). \quad (8.6)$$

- K **strongly augmented views** $\{(I_{s_k}^A, I_{s_k}^B)\}_{k=1}^K$ processed by the student model f_{θ} . These views include standard photometric distortions (color jittering, Gaussian blur) and the proposed **ChangeMix** augmentation. Typically, $K = 1$ for FixMatch-style frameworks and $K = 2$ for UniMatch-style frameworks.

The unsupervised loss $\mathcal{L}_{unsup}$ enforces that the student model's predictions on the strongly augmented (and ChangeMixed) views match the teacher's pseudo-label $\hat{M}$, filtered by a confidence threshold τ (set to 0.95 in all experiments):

$$\mathcal{L}_{unsup} = \frac{1}{K} \sum_{k=1}^K \mathcal{L}_{CE}(f_{\theta}(I_{s_k}^A, I_{s_k}^B), \hat{M}) \cdot \mathbf{1}[\hat{C} \geq \tau], \quad (8.7)$$

where $\mathcal{L}_{CE}$ is the standard cross-entropy loss. By integrating ChangeMix into the strong augmentation pipeline, the model is consistently trained on synthetic "change" scenarios derived from high-confidence predictions.

Training Objective

The total loss function combines supervised and unsupervised components:

$$\mathcal{L} = \frac{1}{2} \mathcal{L}_{sup} + \frac{1}{2} \mathcal{L}_{unsup}, \quad (8.8)$$

The supervised loss $\mathcal{L}_{sup}$ is the standard cross-entropy between the student model's predictions on weakly augmented labeled pairs and the ground-truth masks:

$$\mathcal{L}_{sup} = \mathcal{L}_{CE}(f_{\theta}(I_w^A, I_w^B), M). \quad (8.9)$$

8.3.3 ChangeMix: Change-Aware Cut-and-Paste Augmentation

ChangeMix functions as a bitemporal, instance-level augmentation for SSCD. It increases the effective frequency and diversity of changed regions while preserving the weak-to-strong consistency used in teacher-student frameworks. Within this pipeline, ChangeMix injects a targeted supervisory signal for the minority "changed" class by pasting paired pre-change and post-change instances into strongly augmented unlabeled views. Unlike generic region mixing techniques (such as the paired CutMix embedded in UniMatch V2 or Copy-Paste) that sample regions regardless of semantic rarity, ChangeMix maintains a change-aware instance bank and explicitly balances the temporal bias (appearance versus disappearance) during pasting. The augmentation operates in two complementary stages at each iteration: instance extraction into the Instance Bank, and instance pasting into strongly augmented unlabeled pairs. Figure 8.2 illustrates the full pipeline.

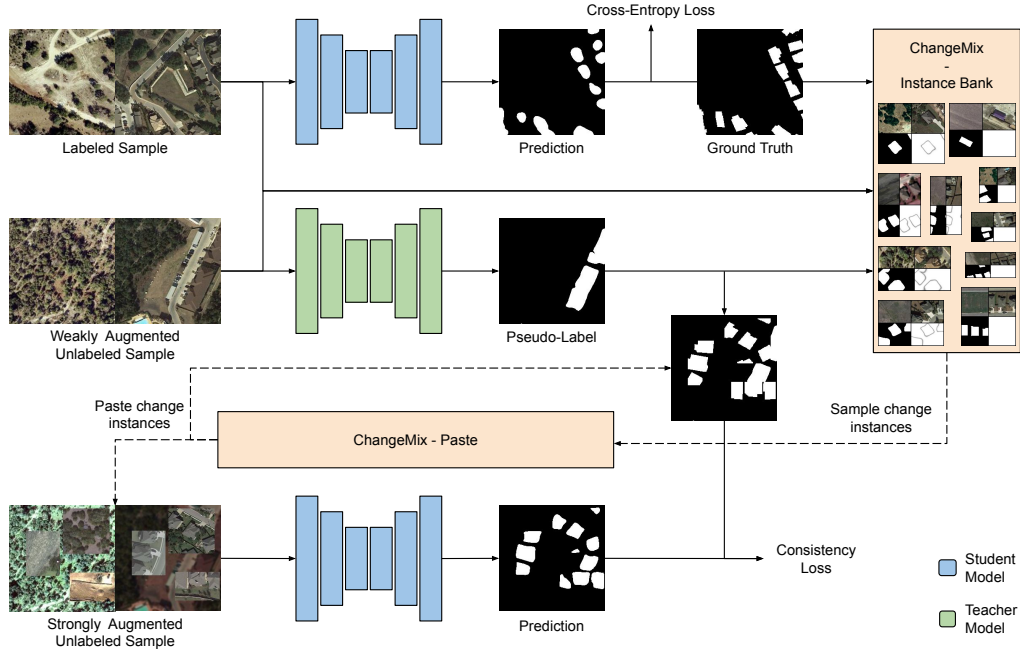


Figure 8.2. Overview of the proposed ChangeMix within a teacher-student semi-supervised change detection pipeline. The teacher model predicts on a weakly augmented view of unlabeled pairs to produce pseudo-labels and confidence maps. High-confidence change instances together with ground-truth change instances are collected in an Instance Bank as paired pre- and post-change crops along with their masks and confidences. ChangeMix then samples multiple instances and pastes them into strongly augmented unlabeled views, updating the corresponding pseudo-labels and confidence maps. The student model is optimized with a consistency loss on these mixed strong views alongside the supervised loss on labeled data.

Instance Extraction and Instance Bank

At each training step, connected components are extracted from a binary change map $M \in \{0, 1\}^{H \times W}$ and a confidence map $C \in [0, 1]^{H \times W}$. These components are sourced from both labeled batches (with confidence strictly equal to 1) and unlabeled batches via the teacher model’s pseudo-labels on the weakly augmented view. This approach leverages scarce ground truth while progressively distilling reliable changes from unlabeled data.

A connected components analysis on $M = 1$ is executed to obtain object-level regions and extract tight bounding boxes. An instance is accepted if its bounding box dimensions fall within $[s_{\min}, s_{\max}]$ and its mean confidence over the foreground pixels lies in $[\tau_{\min}, \tau_{\max}]$. The parameter $\tau_{\min}$ filters out noisy, unreliable predictions to maintain bank quality, while $\tau_{\max}$ excludes overly confident samples to prevent the model from overfitting to them. The crops are padded with p pixels, with the padding explicitly clipped to valid image boundaries. Accepted instances are added to the Instance Bank as tuples (x^A, x^B, m, c) : paired pre-change and post-change crops, together with the corresponding binary change map and confidence map. The bank has a fixed capacity B and follows a First-In-First-Out (FIFO) replacement policy, allowing the set of pasteable instances to track the teacher’s improvements

during training. To increase diversity, instance extraction can utilize either weakly or strongly augmented copies of the current batch. This practice ensures that the Instance Bank remains populated with instances consistent with the augmentation space processed by the student model.

Instance pasting

The ChangeMix pasting operation is applied to an unlabeled batch with probability $p_{mix} = 0.5$. The algorithm samples up to N instances from the bank. For each instance, up to L random placements are attempted to find a suitable location that does not overlap with previously pasted regions. The paired crops (x^A, x^B) are then pasted into the corresponding windows of the strongly augmented view.

To mitigate temporal bias, the temporal order of pre-change and post-change patches is randomly swapped before pasting with probability $p_{swap} = 0.5$. In this way, appearance and disappearance events are simulated with equal probability. Finally, the initial teacher pseudo-label and confidence maps are updated with the instance’s m and c maps, ensuring the supervisory signal is consistent with the newly injected changes. Through this mechanism, every successful paste operation directly raises the effective sampling rate of the minority class without destabilizing the training process.

8.4 Experiments

8.4.1 Datasets

The effectiveness of ChangeMix is evaluated on two very-high-resolution (VHR) remote sensing benchmarks for building change detection, namely LEVIR-CD [20], and WHU-CD [63].

LEVIR-CD [20] is a large-scale dataset containing 637 pairs of 1024×1024 pixels Google Earth images, with a spatial resolution of 0.5 m/pixel. The images were captured over 20 different regions in Texas, USA, between 2002 and 2018, and contain over 31,000 instances of building changes. The official data splits for training, validation, and testing are maintained throughout the evaluation.

WHU-CD [63] consists of a pair of large VHR (0.3 m/pixel) aerial images covering a 20.5 km^2 area in Christchurch, New Zealand, that was rebuilt after an earthquake. The bi-temporal images, sized $32,507 \times 15,354$ pixels, document significant building changes between 2012 and 2016. Following common practice, the images are cropped into non-overlapping 256×256 tiles. The official training, validation, and testing splits are utilized.

For both datasets, a semi-supervised learning scenario is simulated by utilizing only a small fraction of the official training set labels for direct supervision, treating the remainder as unlabeled data. To ensure a fair and reproducible comparison, for each label budget (5%, 10%, 20%, 40%), the labeled and unlabeled subsets are composed of the exact image tile IDs used in previous studies such as UniMatch [162] and UniMatch V2 [163].

8.4.2 Evaluation Metrics

Performance is assessed by treating CD as a binary segmentation task where "change" is the positive class. The primary metric is the Intersection-over-Union for the changed class (IoU_c), as it provides a robust measure of accuracy for the minority class of interest. For a comprehensive comparison with existing literature, Overall Accuracy (OA) is also reported.

For all experiments involving the teacher-student frameworks FixMatch and UniMatch V2, the EMA (teacher) model that achieves the highest IoU_c on the validation set is saved for testing. For the UniMatch V1, "Labeled Only", and "Fully Supervised" experiments, the standard model is saved using the same criterion. All final results are reported on the test set.

8.4.3 Implementation Details

The implementation is built upon the official PyTorch-based UniMatch V2 [163] framework, which was adapted for the comparative experiments presented in this study.

Network Architectures

The primary experiments employ a Dense Prediction Transformer (DPT) [120] architecture, using DINOv2 [106] backbones: DINOv2-Small (ViT-S/14) and DINOv2-Base (ViT-B/14). These result in models with 24.8M and 97.5M parameters, respectively. To validate the model-agnostic nature of ChangeMix, a convolutional architecture is also included. For this purpose, the DeepLabV3+ [21] implementation from UniMatch V1 was adapted to the UniMatch V2 framework, using a ResNet-50 backbone (40.5M parameters).

SSL Frameworks

To demonstrate that ChangeMix is framework-agnostic, it is integrated into three distinct SSL frameworks: FixMatch [134], UniMatch V1 [162], and UniMatch V2 [163]. For these experiments, the DPT architecture with DINOv2-Base was utilized. This required adapting the official UniMatch V1 framework, which was not originally designed for this architecture, to ensure a controlled comparison. In all baseline experiments, standard CutMix is used among the strong augmentations, which is subsequently replaced with ChangeMix.

Training Configuration

All models were trained for 60 epochs with a batch size of 8. Experiments were executed on either a single NVIDIA GeForce RTX 3090 GPU or a single NVIDIA A100 GPU, depending on resource availability. Different training parameters were employed depending on the architecture. For the DPT architecture with DINOv2, an AdamW optimizer was used with a learning rate of $5e-6$ for the encoder and $2e-4$ for the decoder. The weight decay was set to 0.01. For DeepLabV3+ with ResNet-50, an SGD optimizer was used with a learning rate of 0.02, momentum of

0.9, and weight decay of $1e-4$. For all configurations, the learning rates followed a polynomial decay schedule with a power of 0.9.

Data Augmentation

Weak augmentations consist of random resizing, random cropping, and horizontal flipping. Strong augmentations include the same geometric transformations along with color jittering, Gaussian blurring, and either the baseline CutMix or the proposed ChangeMix.

ChangeMix Parameters

The Instance Bank capacity is set to $B = 500$. The augmentation is applied to an unlabeled batch with a probability of $p_{mix} = 0.5$. During pasting, up to $N = 10$ instances are pasted, with $L = 10$ placement attempts per instance to prevent spatial overlap. To mitigate temporal bias, the pre-change and post-change patches are temporally swapped with a probability of $p_{swap} = 0.5$ prior to pasting. For instance extraction, only regions with bounding box side lengths between $s_{min} = 8$ and $s_{max} = 192$ pixels are accepted. Extracted instances are padded with $p = 32$ pixels to provide spatial context. Finally, the acceptable confidence range for pseudo-labeled instances is bounded at $[\tau_{min}, \tau_{max}] = [0.95, 0.99]$, whereas ground-truth instances are invariably accepted with a confidence score of 1. The sensitivity of the model to the primary hyperparameters is analyzed in Section 8.4.6.

8.4.4 Comparison Experiments

A comprehensive set of experiments is conducted to validate ChangeMix and to test two primary hypotheses: (1) ChangeMix is *framework-agnostic*, improving performance across different SSL methods, and (2) ChangeMix is *model-agnostic*, providing gains across different network architectures.

Framework-Agnosticism

ChangeMix is first compared against its baseline (standard CutMix) within three SSL frameworks: FixMatch [134], UniMatch V1 [162], and UniMatch V2 [163]. All experiments in this section use the DPT architecture with DINOv2-Base for a controlled comparison. The results are summarized in Table 8.1 (LEVIR-CD) and Table 8.2 (WHU-CD), alongside other published SSCD methods for context.

On the LEVIR-CD dataset (Table 8.1), substituting baseline CutMix with ChangeMix yields consistent IoU_c improvements across all frameworks and label budgets. The gains range from +0.23 to +1.22 points. This improvement is particularly notable at the 5% label budget for UniMatch V2, where performance increases from 82.62 to 83.83 (+1.21 points). Similarly, at the 10% label budget, ChangeMix boosts IoU_c from 83.43 to 84.65 (+1.22) for FixMatch and from 83.33 to 84.49 (+1.16) for UniMatch V2. Moreover, at 40% label budget, ChangeMix elevates the performance of all frameworks to levels that approach or match the fully supervised baseline. Notably, UniMatch V1 with ChangeMix reaches an IoU_c of 85.07, effectively matching the fully supervised IoU_c of 85.06. This result highlights the

specific nature of the annotation bottleneck in this domain. Matching fully supervised performance with only a fraction of the labels occurs because the primary limiting factor for the baseline models is the scarcity of the "changed" class, rather than a lack of overall image context. By actively pasting multiple high-confidence change instances into the training batches, ChangeMix increases the effective sampling rate of these rare events. At a 40% label budget, the combination of the available ground truth and this densified pseudo-supervision provides a learning signal for the minority class that is comparable to the 100% labeled dataset.

Table 8.1. Comparison of semi-supervised change detection methods on the LEVIR-CD test set. Results are reported as changed-class intersection-over-union (IoU_c) and overall accuracy (OA).

Method	Backbone	LEVIR-CD							
		5% (356)		10% (712)		20% (1424)		40% (2848)	
		IoU_c	OA	IoU_c	OA	IoU_c	OA	IoU_c	OA
AdvEnt [147]	FCN-RN-50	66.1	98.08	72.3	98.45	74.6	98.58	75.0	98.60
S4GAN [98]	FCN-RN-50	64.0	97.89	67.0	98.11	73.4	98.51	75.4	98.62
SemiCDNet [113]	UNet++	67.6	98.17	71.5	98.42	74.3	98.58	75.5	98.63
SemiCD [6]	FCN-RN-50	72.5	98.47	75.5	98.63	76.2	98.68	77.2	98.72
CutMix-CD [131]	FCN-RN-50	78.2	98.78	79.8	98.87	80.9	98.95	-	-
UniMatch V1 [162]	PSPNet-RN-50	75.6	98.62	79.0	98.83	79.0	98.84	78.2	98.79
UniMatch V1 [162]	DeepLab-RN-50	80.7	98.95	82.0	99.02	81.7	99.02	82.1	99.03
SemiCD-VL [85]	RN-50 + VLM	81.9	99.02	82.6	99.06	82.7	99.05	83.0	99.07
Labeled Only	DINOv2-B	78.95	98.82	81.55	98.98	81.95	99.00	83.39	99.08
FixMatch	DINOv2-B	82.77	99.05	83.43	99.09	83.80	99.12	84.10	99.13
+ ChangeMix	DINOv2-B	83.63	99.11	84.65	99.16	84.32	99.15	84.90	99.18
UniMatch V1	DINOv2-B	83.30	99.09	83.89	99.12	84.04	99.13	84.61	99.15
+ ChangeMix	DINOv2-B	83.53	99.10	84.56	99.16	84.86	99.17	85.07	99.18
UniMatch V2	DINOv2-B	82.62	99.05	83.33	99.09	84.11	99.13	84.25	99.14
+ ChangeMix	DINOv2-B	83.83	99.12	84.49	99.16	84.81	99.17	84.93	99.18
Fully Supervised	DINOv2-B	100% (7120)		IoU_c : 85.06		OA: 99.18			

Similar trends are observed on the WHU-CD dataset (Table 8.2). ChangeMix consistently improves upon the baseline frameworks, and the benefits are evident across all data regimes. For example, at the 20% label budget, ChangeMix increases the IoU_c of FixMatch by +2.88 points (from 85.05 to 87.93).

In the very low-data setting of 5% labels, the baseline models exhibit significant instability and often converge to suboptimal solutions. This instability arises because the models become overly confident too rapidly due to the overwhelming prevalence of background pixels in the sparse labeled set. Consequently, they rely prematurely and excessively on the unlabeled branch, becoming trapped in a poor-performance state driven by confirmation bias. This dynamic is corroborated by the "Labeled Only" experiment, which achieves a higher IoU_c of 68.91 at the 5% budget, significantly outperforming the baseline FixMatch (28.30) and UniMatch V2 (27.03) models.

Furthermore, the UniMatch V1 baseline reaches higher performance (62.73) in this regime specifically because it lacks an EMA teacher. Without the EMA mechanism, the teacher's weights oscillate more freely during training, occasionally navigating toward viable parameter spaces. In contrast, the EMA-based teachers

in FixMatch and UniMatch V2 stabilize prematurely in low-performance regimes, predicting only the most certain, often background, regions. By consistently injecting verified change instances into the training loop, ChangeMix counteracts this early confirmation bias, providing a stabilizing effect that improves the IoU_c by +15.09 points for FixMatch and +9.68 points for UniMatch V2 at the 5% label budget.

However, it must be acknowledged that the absolute performance of these models at the 5% budget, even with ChangeMix (43.39 and 36.71 IoU_c , respectively), remains distinctly poor compared to the labeled-only baseline. This indicates that while the augmentation provides a necessary signal boost and partially mitigates the early confirmation bias, it cannot entirely overcome the structural flaw of the SSL frameworks relying too heavily on the unlabeled data during the initial training phase. Resolving this fundamental instability would likely require architectural adjustments, such as implementing a ramp-up schedule or postponing the introduction of the unsupervised loss, which remains outside the scope of this augmentation-focused study.

Table 8.2. Comparison of semi-supervised change detection methods on the WHU-CD test set. Results are reported as changed-class intersection-over-union (IoU_c) and overall accuracy (OA).

Method	Backbone	WHU-CD							
		5% (297)		10% (594)		20% (1189)		40% (2378)	
		IoU_c	OA	IoU_c	OA	IoU_c	OA	IoU_c	OA
AdvEnt [147]	FCN-RN-50	55.1	-	61.6	-	73.8	98.80	76.6	98.94
S4GAN [98]	FCN-RN-50	18.3	96.69	62.6	98.15	70.8	98.60	76.4	98.96
SemiCDNet [113]	UNet++	51.7	-	62.0	-	66.7	98.28	75.9	98.93
SemiCD [6]	FCN-RN-50	65.8	98.37	68.1	98.47	74.8	98.84	77.2	98.96
UniMatch V1 [162]	PSPNet-RN-50	77.5	99.06	78.9	99.10	82.9	99.26	84.4	99.32
UniMatch V1 [162]	DeepLab-RN-50	80.2	99.15	81.7	99.22	81.7	99.18	85.1	99.35
SemiCD-VL [85]	RN-50 + VLM	81.8	-	83.2	-	84.8	99.36	85.7	99.39
Labeled Only	DINOv2-B	68.91	98.65	75.49	98.92	80.40	99.13	86.29	99.42
FixMatch	DINOv2-B	28.30	97.09	80.45	99.19	85.05	99.38	87.77	99.48
+ ChangeMix	DINOv2-B	43.39	97.68	81.75	99.24	87.93	99.49	88.90	99.53
UniMatch V1	DINOv2-B	62.73	98.42	77.90	99.09	85.15	99.36	86.88	99.45
+ ChangeMix	DINOv2-B	76.82	99.05	79.15	99.12	85.93	99.41	87.60	99.48
UniMatch V2	DINOv2-B	27.03	97.05	80.89	99.20	86.83	99.45	87.65	99.48
+ ChangeMix	DINOv2-B	36.71	97.43	81.95	99.24	87.88	99.50	88.06	99.50
Fully Supervised	DINOv2-B	100% (5947)		IoU_c : 89.03		OA: 99.54			

Collectively, these results confirm that the change-aware augmentation strategy, which explicitly targets the minority class, is more effective than generic, non-targeted mixing augmentations like CutMix.

Model-Agnosticism

To demonstrate that ChangeMix is not limited to a specific model, it is evaluated within the UniMatch V2 framework using three different network configurations: DPT with DINOv2-Small, DPT with DINOv2-Base, and DeepLabV3+ with ResNet-50. Results are presented in Table 8.3 (LEVIR-CD) and Table 8.4 (WHU-CD).

Across all architectures and both datasets, ChangeMix consistently outperforms the baseline. The improvements are especially pronounced in the DeepLabV3+ architecture, which struggles with the baseline UniMatch V2 setup. For example, on LEVIR-CD at 20% labels, ChangeMix adds +1.01 IoU_c for DINOv2-Small, +0.70 IoU_c for DINOv2-Base, and a substantial +5.73 IoU_c for DeepLabV3+. This confirms that ChangeMix functions as a robust, model-agnostic augmentation that provides significant value across varying network capacities.

Table 8.3. Backbone comparison for UniMatch v2 and the ChangeMix augmentation on the **LEVIR-CD** test set.

Method	Backbone	LEVIR-CD							
		5% (356)		10% (712)		20% (1424)		40% (2848)	
		IoU _c	OA	IoU _c	OA	IoU _c	OA	IoU _c	OA
Labeled Only	DINOv2-S	78.38	98.79	80.23	98.90	81.08	98.95	82.19	99.02
UniMatch V2	DINOv2-S	81.77	99.00	82.77	99.05	83.12	99.07	83.18	99.08
+ ChangeMix	DINOv2-S	82.80	99.06	83.47	99.10	84.13	99.13	84.27	99.14
Fully Supervised	DINOv2-S	100% (7120)		IoU _c : 84.16		OA: 99.14			
Labeled Only	DINOv2-B	78.95	98.82	81.55	98.98	81.95	99.00	83.39	99.08
UniMatch V2	DINOv2-B	82.62	99.05	83.33	99.09	84.11	99.13	84.25	99.14
+ ChangeMix	DINOv2-B	83.83	99.12	84.49	99.16	84.81	99.17	84.93	99.18
Fully Supervised	DINOv2-B	100% (7120)		IoU _c : 85.06		OA: 99.18			
Labeled Only	DeepLabV3+	72.96	98.53	77.39	98.77	78.42	98.82	79.31	98.87
UniMatch V2	DeepLabV3+	75.31	98.68	76.69	98.76	76.37	98.75	77.78	98.82
+ ChangeMix	DeepLabV3+	78.03	98.83	80.47	98.96	82.10	99.04	82.07	99.04
Fully Supervised	DeepLabV3+	100% (7120)		IoU _c : 81.45		OA: 99.01			

8.4.5 Ablation Experiments

To validate the contribution of each component within ChangeMix, a progressive ablation study was conducted on the LEVIR-CD dataset with a 20% label budget, using the UniMatch V2 framework and the DPT architecture with DINOv2-Small. Results are presented in Table 8.5. Starting from the baseline with standard CutMix (83.12 IoU_c), introducing change-aware instance selection from the current batch provides an initial gain of +0.30 points. Integrating the Instance Bank, which accumulates high-quality examples over time, further increases performance by +0.13 points. Enabling the pasting of multiple instances per image (up to 10) raises the IoU_c by an additional +0.32 points to 83.87. Finally, incorporating random temporal swapping of pre-change and post-change patches to mitigate temporal bias provides a final +0.26 improvement, reaching 84.13 IoU_c. This demonstrates that each component contributes meaningfully to the overall performance, yielding a cumulative +1.01 point gain over the baseline.

8.4.6 Hyperparameter Experiments

The sensitivity of ChangeMix to its key hyperparameters (bank capacity, maximum instance size, confidence thresholds, and padding) is analyzed across four settings using the 20% label budget, which we refer to as: LEVIR-Base (LEVIR-CD with

Table 8.4. Backbone comparison for UniMatch v2 and the ChangeMix augmentation on the **WHU-CD** test set.

Method	Backbone	WHU-CD							
		5% (297)		10% (594)		20% (1189)		40% (2378)	
		IoU _c	OA	IoU _c	OA	IoU _c	OA	IoU _c	OA
Labeled Only	DINOv2-S	65.30	98.40	68.80	98.58	76.40	98.91	81.30	99.19
UniMatch V2	DINOv2-S	19.60	96.71	80.14	99.15	84.77	99.36	85.44	99.38
+ ChangeMix	DINOv2-S	34.80	97.35	80.74	99.19	86.99	99.45	86.36	99.41
Fully Supervised	DINOv2-S	100% (5947)		IoU _c : 88.71		OA: 99.53			
Labeled Only	DINOv2-B	68.91	98.65	75.49	98.92	80.40	99.13	86.29	99.42
UniMatch V2	DINOv2-B	27.03	97.05	80.89	99.20	86.83	99.45	87.65	99.48
+ ChangeMix	DINOv2-B	36.71	97.43	81.95	99.24	87.88	99.50	88.06	99.50
Fully Supervised	DINOv2-B	100% (5947)		IoU _c : 89.03		OA: 99.54			
Labeled Only	DeepLabV3+	58.52	98.10	64.73	98.26	73.65	98.74	79.80	99.09
UniMatch V2	DeepLabV3+	56.86	97.76	58.07	97.45	68.37	98.30	84.68	99.34
+ ChangeMix	DeepLabV3+	66.24	98.60	79.43	99.09	83.31	99.26	86.28	99.42
Fully Supervised	DeepLabV3+	100% (5947)		IoU _c : 87.03		OA: 99.45			

Table 8.5. Ablation study on the components of ChangeMix. Each row represents the addition of a new component to the configuration of the row above it. All experiments are conducted on the LEVIR-CD dataset with a 20% label budget using the DPT architecture with DINOv2-Small.

Method	IoU _c	OA	F1
Labeled Only	81.08	98.95	89.55
UniMatch V2 (with baseline CutMix)	83.12	99.07	90.78
+ Change Aware (from current iteration)	83.42	99.09	90.96
+ Instance Bank	83.55	99.10	91.04
+ Multiple Instances	83.87	99.12	91.23
+ Temporal Swap (full ChangeMix)	84.13	99.13	91.38

DINOv2-Base, Table 8.6), LEVIR-Small (LEVIR-CD with DINOv2-Small, Table 8.6), WHU-Base (WHU-CD with DINOv2-Base, Table 8.7), and WHU-Small (WHU-CD with DINOv2-Small, Table 8.7). A single optimal configuration emerged across all testbeds.

- **Bank Size (500):** This parameter governs the memory-quality tradeoff. A bank that is too small (20) limits diversity and risks overfitting to a few repeated instances. Conversely, a significantly larger capacity (2500) acts as a long-term memory that retains stale, lower-quality pseudo-labels generated during the early epochs of training. A capacity of 500 ensures the bank continuously refreshes with the teacher model’s most recent and accurate predictions.
- **Max Instance Size (192):** This parameter acts as an admission filter for the Instance Bank, dictating the scale of the objects available for augmentation. At the VHR resolutions of these datasets (0.3 to 0.5 meters per pixel), 192

pixels correspond to approximately 60 to 90 meters. This limit is sufficiently large to admit the majority of standard residential and commercial buildings into the bank. Smaller limits (e.g., 128) result in the outright rejection of larger structures during the extraction phase, depriving the model of targeted supervision for those specific instances. Conversely, larger limits (e.g., 256) admit oversized regions that disrupt the multi-instance pasting strategy. If an extremely large instance happens to be pasted first, it overtakes the majority of the augmented view; if it is selected later, it almost invariably fails the non-overlapping placement constraints and cannot be pasted at all.

- **Confidence Thresholds [0.95–0.99]:** This range effectively mitigates confirmation bias and selects the most useful pseudo-labels. The 0.95 minimum filters out noisy predictions, while capping the maximum at 0.99 prevents the network from exclusively sampling well-learned examples that the model already predicts with very high confidence. By forcing the Instance Bank to collect high-quality but slightly more challenging instances, the augmentation continuously pushes the student model to refine its decision boundaries.
- **Padding (32):** Pasting an object with zero padding creates abrupt, unnatural pixel boundaries that neural networks can easily identify as artificial artifacts. Adding 32 pixels of padding includes the natural shadow, immediate terrain, and edges of the building. This provides the network with the necessary spatial context to learn realistic change boundaries, without importing so much background that it defeats the targeted nature of the augmentation.

Based on this comprehensive analysis, the configuration consisting of a Bank Size of 500, a Max Instance size of 192, a Confidence range of [0.95–0.99], and a Padding of 32 proved to be optimal across the evaluated benchmarks. It is important to note that this specific configuration was optimal for both LEVIR-CD and WHU-CD largely due to their similar characteristics and shared focus on Very High Resolution building change detection. Deploying ChangeMix on datasets with drastically different scales, target subjects (such as deforestation or agricultural crop changes), or input image sizes would likely require recalibrating these hyperparameters to maintain optimal performance. Nevertheless, to ensure a fair and rigorous evaluation, this single configuration was adopted for all experiments presented in Tables 8.1, 8.2, 8.3, and 8.4, without dataset or backbone-specific finetuning. These ablations confirm that while our hyperparameters are well-tuned, the method is robust, and even with suboptimal parameters, ChangeMix consistently provides a significant improvement over the baseline.

8.4.7 Qualitative Results

Qualitative comparisons are provided to visually demonstrate the advantages of the proposed method over the baseline. Figure 8.3 presents examples from the LEVIR-CD test set, while Figure 8.4 displays examples from the WHU-CD test set. These examples are drawn from the 10% label budget experiments using the DPT architecture with DINOv2-Base.

Table 8.6. Instance Bank hyperparameter evaluation on LEVIR-CD (20% labels) using the DPT architecture with DINOv2-Base and DINOv2-Small.

LEVIR-CD (20% labels)					
Parameter	Value	DINOv2-B		DINOv2-S	
		IoU _c	OA	IoU _c	OA
Bank Size	20	84.65	99.16	83.41	99.09
	100	84.65	99.16	83.80	99.12
	500	84.81	99.17	84.13	99.13
	2500	84.73	99.17	83.48	99.10
Max Instance Size	128	84.74	99.17	83.71	99.11
	160	84.58	99.16	83.98	99.13
	192	84.81	99.17	84.13	99.13
	256	84.51	99.15	83.57	99.10
Min Confidence	0.925	84.78	99.17	83.93	99.12
	0.95	84.81	99.17	84.13	99.13
	0.975	84.73	99.17	83.73	99.11
Max Confidence	0.975	84.64	99.16	83.83	99.12
	0.99	84.81	99.17	84.13	99.13
	1	84.67	99.17	83.90	99.12
Padding	0	84.35	99.15	83.41	99.09
	4	84.47	99.15	83.47	99.10
	16	84.67	99.16	83.71	99.11
	32	84.81	99.17	84.13	99.13
	64	84.76	99.17	83.80	99.11

In Figure 8.3, the baseline model exhibits limitations that ChangeMix resolves. In Sample 1, the proposed model produces a more complete mask for the structure, whereas the baseline prediction is fragmented. Samples 3 and 4 contain instances of new building constructions (appearance events) that the baseline fails to detect. Moreover, Samples 2 and 4 provide visual evidence of temporal bias mitigation. Both scenes feature disappearance events that the baseline model misses, whereas the model using ChangeMix, which benefits from temporal swapping, correctly identifies these changes. Sample 4 is a complex example including both an appearance and a disappearance event, and the proposed model successfully captures both.

Similar trends are observed on the WHU-CD dataset in Figure 8.4. The baseline misses significant portions of the change (Samples 2 through 4) or entire new buildings (Samples 1 and 2), while the proposed method generates more complete and accurate predictions. Interestingly, Sample 2 shows an example of a correct prediction by both models where the ground truth is incomplete. This type of label noise may contribute to the instability observed in the models in the very-low data regime (5% labels). These examples visually corroborate the quantitative findings, demonstrating that ChangeMix reduces false negatives and produces masks that are more faithful to the ground truth.

Table 8.7. Instance Bank hyperparameter evaluation on WHU-CD (20% labels) using the DPT architecture with DINOv2-Base and DINOv2-Small.

WHU-CD (20% labels)					
Parameter	Value	DINOv2-B		DINOv2-S	
		IoU _c	OA	IoU _c	OA
Bank Size	20	87.58	99.48	86.17	99.41
	100	87.00	99.45	86.52	99.44
	500	87.88	99.50	86.99	99.45
	2500	87.75	99.49	85.46	99.38
Max Instance Size	128	87.15	99.46	85.83	99.40
	160	87.37	99.47	86.34	99.43
	192	87.88	99.50	86.99	99.45
	256	87.08	99.47	86.65	99.43
Min Confidence	0.925	87.58	99.48	85.09	99.36
	0.95	87.88	99.50	86.99	99.45
	0.975	86.64	99.44	85.36	99.38
Max Confidence	0.975	87.44	99.47	86.11	99.41
	0.99	87.88	99.50	86.99	99.45
	1	86.38	99.43	85.66	99.39
Padding	0	86.72	99.45	85.97	99.40
	4	87.39	99.47	84.00	99.30
	16	87.79	99.49	86.16	99.41
	32	87.88	99.50	86.99	99.45
	64	87.13	99.47	83.49	99.29

8.5 Concluding Remarks

This chapter presented ChangeMix, a change-aware cut-and-paste augmentation strategy for semi-supervised remote sensing change detection. The proposed approach addresses three fundamental challenges in this domain: the scarcity of pixel-level annotations, the pronounced class imbalance between changed and unchanged pixels, and the temporal bias inherent in detecting appearances versus disappearances.

At the core of ChangeMix is an Instance Bank mechanism, which dynamically accumulates change instances from both limited labeled data and high-confidence pseudo-labels obtained from the abundant unlabeled data. Multiple instances, in the form of paired pre-change and post-change crops with their corresponding masks and confidences, are sampled and pasted onto strongly augmented samples within a teacher-student framework. This process effectively increases the sampling frequency of the minority "changed" class, providing the targeted supervision necessary to improve change detection performance. Furthermore, by randomly swapping the temporal order of pre-change and post-change patches before pasting, ChangeMix simulates appearance and disappearance events with equal probability, mitigating temporal bias.

A comprehensive validation confirmed the efficacy and versatility of the proposed approach. By integrating ChangeMix into three distinct semi-supervised learning pipelines (FixMatch, UniMatch V1, and UniMatch V2) and evaluating it across multiple architectures (DPT with DINOv2 and DeepLabV3+), the method demon-

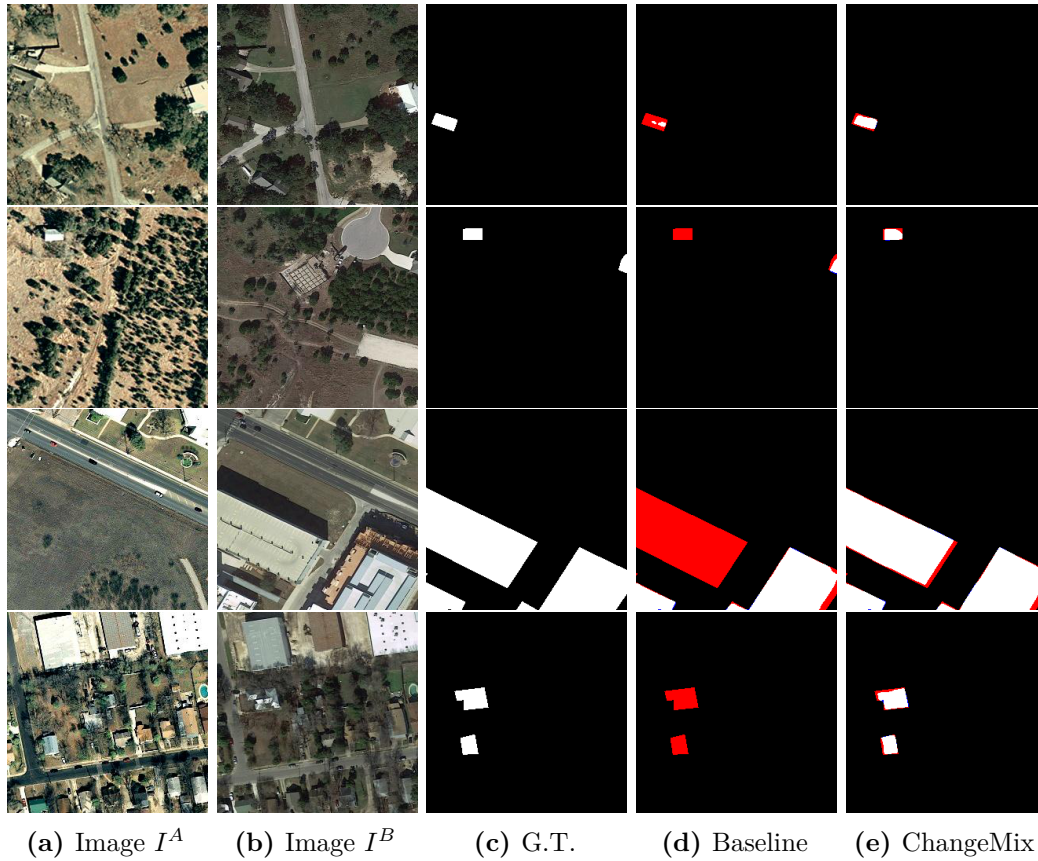


Figure 8.3. Qualitative comparison of change detection results on the LEVIR-CD test set using a 10% label budget and the DPT architecture with a DINOv2-Base backbone. Columns denote: (a) Pre-change Image I^A , (b) Post-change Image I^B , (c) Ground Truth (G.T.), (d) Baseline (UniMatch V2), and (e) Proposed Method (UniMatch V2 + ChangeMix). Compared to the baseline, the proposed ChangeMix augmentation produces more complete masks and mitigates temporal bias by accurately capturing both appearance (constructions) and disappearance (demolitions) events. Legend: **Red** = False Negatives, **Blue** = False Positives, White = True Positives.

strated consistent improvements across all tested configurations and label budgets on the LEVIR-CD and WHU-CD benchmarks. Ablation studies further confirmed that each component of the proposed method contributes meaningfully to the overall performance gains.

Ultimately, ChangeMix serves as a practical, framework-agnostic augmentation that seamlessly complements existing semi-supervised pipelines. By demonstrating that the semantic-aware recombination of available data can effectively compensate for a lack of manual annotations, this chapter reinforces the central thesis that compositional synthesis is a robust and transferrable strategy for overcoming the annotation bottleneck in complex perception tasks.

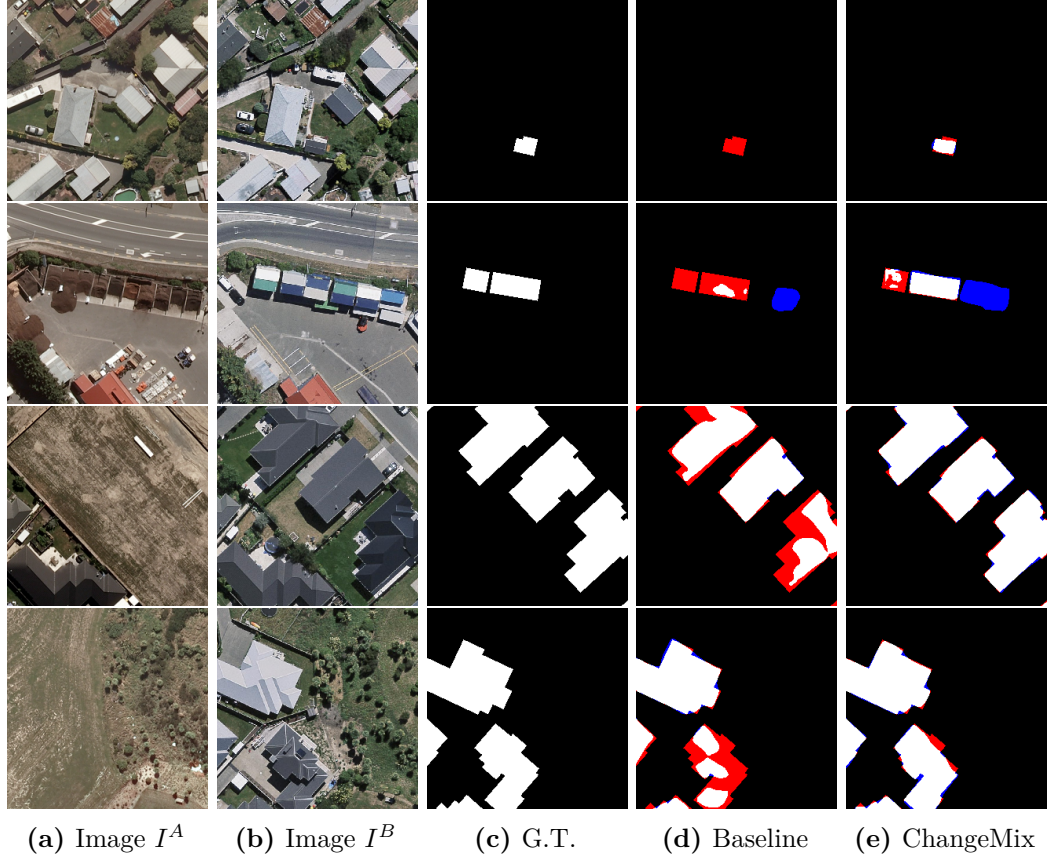


Figure 8.4. Qualitative comparison of change detection results on the WHU-CD test set using a 10% label budget and the DPT architecture with a DINOv2-Base backbone. Columns denote: (a) Pre-change Image I^A , (b) Post-change Image I^B , (c) Ground Truth (G.T.), (d) Baseline (UniMatch V2), and (e) Proposed Method (UniMatch V2 + ChangeMix). The proposed method significantly reduces false negatives, successfully recovering entire missing buildings and restoring significant portions of structural boundaries that the baseline fails to detect. Legend: Red = False Negatives, Blue = False Positives, White = True Positives.

Chapter 9

Conclusion

The deployment of deep learning models in physical, unstructured domains like precision agriculture and remote sensing is frequently limited by the annotation bottleneck. Acquiring high-quality, dense annotations is an expensive and time-consuming process, and this difficulty is often compounded by persistent covariate shifts, extreme class imbalances, and spatially sparse targets. This dissertation presented a cohesive methodological approach to mitigate data scarcity by focusing on label-efficient learning strategies. The research explored two primary mechanisms: surrogate supervision and compositional synthesis. By extracting structured training targets from weak cues and dynamically recombining object-level instances, this work demonstrated how to maximize the utility of available data and generate targeted synthetic supervision.

9.1 Summary of Contributions

The specific contributions of this research were validated across two distinct but methodologically related application domains.

In the first part of the thesis, the focus was on precision agriculture, specifically automated applications in table grape vineyards where environmental variability is prominent.

Chapter 3 introduced a weakly and semi-supervised pipeline to address domain shifts in object detection and instance segmentation. By combining automated bounding box propagation across video frames with classical computer vision refinement techniques like GrabCut, the system generated high-quality pseudo-labels. This approach significantly reduced the reliance on manual pixel-level annotation while adapting source-trained models to specific target fields.

Chapter 4 extended the concept of data amplification by introducing a hybrid data generation pipeline. Real fruit instances, automatically extracted via foundation models, were contextually blended into simulated 3D environments. This compositional synthesis approach expanded the diversity of the training distribution, improving detection robustness against illumination and viewpoint changes without the need for exhaustive physical data collection.

Chapter 5 adapted this compositional synthesis strategy to the complex task of anomaly detection. Given the inherent rarity of agricultural defects, the method

utilized classical edge detection to isolate anomalous surface textures and seamlessly blend them onto healthy fruit. This generated realistic synthetic anomalies, providing a scalable solution to train reliable classifiers for rare events under strictly constrained labeling budgets.

Chapter 6 moved beyond geometric perception to evaluate internal fruit quality. It established that a multi-task deep neural network could reliably estimate Soluble Solid Content using standard, low-cost RGB sensors. By jointly optimizing for visual color classification and continuous sugar content regression, the network effectively captured latent chemical properties. This provided a practical, computationally accessible alternative to specialized and expensive hyperspectral instrumentation.

The second part of the dissertation transferred these core methodological principles to the macroscopic scale of Earth observation.

Chapter 7 adapted instance-level augmentation for satellite semantic segmentation. By extracting discrete components from composite semantic labels, the Cut-and-Paste augmentation technique dynamically reshaped the training distribution. This process effectively increased the representation of minority classes and improved the overall generalization capabilities of the segmentation models.

Chapter 8 addressed the severe class imbalance and temporal bias inherent in remote sensing change detection. The proposed ChangeMix augmentation introduced a dynamic Instance Bank to accumulate high-confidence change regions. By injecting these paired pre-change and post-change instances into unlabeled data and balancing their temporal order, ChangeMix provided targeted supervision for the minority class. The experimental evaluation confirmed that this approach is both framework-agnostic and model-agnostic, yielding consistent performance improvements across various semi-supervised learning configurations.

9.2 Broader Implications

Across both domains, the findings confirm that intelligent data utilization and recombination can effectively compensate for a lack of large-scale manual annotations. Rather than relying exclusively on increasing model capacity or collecting massive datasets, the methods developed in this research emphasize a data-centric approach. By embedding structural priors, leveraging foundation models for zero-shot extraction, and dynamically reshaping the training distribution, perception systems can be made more robust and adaptable to new environments.

This approach lowers the barrier to entry for deploying automated systems in diverse, real-world scenarios. In agriculture, it makes advanced computer vision more accessible for small and medium-sized enterprises that lack the resources to comprehensively re-engineer their fields for automation. In Earth observation, it enables more accurate monitoring of localized, rare geographic events using smaller, specialized datasets.

9.3 Limitations and Future Directions

While the proposed methodologies offer substantial advancements in label-efficient learning, several avenues remain for future exploration.

In the agricultural domain, the methodologies were primarily validated on a specific black table grape variety. Extending these techniques to white grape varieties or entirely different crop types will require further investigation into how subtle visual cues, such as minor hue shifts during ripening, can be reliably captured under varying lighting conditions. Exploring multi-sensor fusion techniques could further improve robustness in more extreme weather environments. Additionally, while compositional synthesis successfully integrates real and simulated environments, future work could explore advanced generative models to further refine the blending realism and automatically generate entirely novel structural anomalies.

In the remote sensing domain, the ChangeMix strategy was validated on high-resolution optical imagery for building change detection. Future research should evaluate its adaptability to medium-resolution satellite data or multi-sensor inputs, such as Synthetic Aperture Radar (SAR). Furthermore, integrating physical constraints or geographical priors into the augmentation pipeline could ensure that the synthetic spatial compositions adhere more strictly to real-world topological rules.

9.4 Concluding Remarks

Ultimately, overcoming the annotation bottleneck requires a shift from passive data collection to active data synthesis and targeted supervision. This dissertation established a comprehensive framework to achieve this, demonstrating that robust perception is attainable even when labeled data is scarce, imbalanced, or difficult to acquire. By providing practical, scalable solutions for both precision agriculture and remote sensing, this research contributes to the broader objective of deploying reliable artificial intelligence in complex scenarios.

La ricerca da cui scaturisce il lavoro di tesi è stata finanziata dall'Unione Europea - Next Generation EU, Missione: I.3.3 Borse PNRR Dottorati innovativi che rispondono ai fabbisogni di innovazione delle imprese (Missione 4), Componente 2, CUP 352: B53C22002020004

Bibliography

- [1] Eu canopies project: A collaborative paradigm for human workers and multi-robot teams in precision agriculture systems. <https://canopies.inf.uniroma3.it/>, 2021.
- [2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.
- [3] Y. Akkem, S. K. Biswas, and A. Varanasi. Smart farming monitoring using ml and mlops. In *International Conference On Innovative Computing And Communication*, pages 665–675. Springer, 2023.
- [4] K. Anderson, B. Ryan, W. Sonntag, A. Kavvada, and L. Friedl. Earth observation in service of the 2030 agenda for sustainable development. *Geospatial Information Science*, 20(2):77–96, 2017.
- [5] C. W. Bac, E. J. van Henten, J. Hemming, and Y. Edan. Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. *Journal of Field Robotics*, 31(6):888–911, 2014.
- [6] W. G. C. Bandara and V. M. Patel. Revisiting consistency regularization for semi-supervised change detection in remote sensing images. *arXiv preprint arXiv:2204.08454*, 2022.
- [7] W. G. C. Bandara and V. M. Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210. IEEE, 2022.
- [8] S. Bang, F. Baek, S. Park, W. Kim, and H. Kim. Image augmentation to improve construction resource detection using generative adversarial networks, cut-and-paste, and image transformation techniques. *Automation in Construction*, 115:103198, 2020.
- [9] R. Barth, J. Hemming, and E. J. van Henten. Improved part segmentation performance by optimising realism of synthetic images using cycle generative adversarial networks. *arXiv preprint arXiv:1803.06301*, 2018.
- [10] R. Barth, J. IJsselmuiden, J. Hemming, and E. V. Henten. Data synthesis methods for semantic segmentation in agriculture: A capsicum annum dataset. *Computers and Electronics in Agriculture*, 144:284–296, 2018.

- [11] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Computer Vision – ECCV 2006*, pages 404–417, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.
- [12] C. Bazinas, E. Vrochidou, T. Kalampokas, A. Karampatea, and V. G. Kaburlasos. A non-destructive method for grape ripeness estimation using intervals’ numbers (ins) techniques. *Agronomy*, 12(7), 2022.
- [13] E. Bellocchio, T. A. Ciarfuglia, G. Costante, and P. Valigi. Weakly supervised fruit counting for yield estimation using spatial consistency. *IEEE Robotics and Automation Letters*, 4(3):2348–2355, 2019.
- [14] J. Bömer, L. Zabawa, P. Sieren, A. Kicherer, L. Klingbeil, U. Rascher, O. Muller, H. Kuhlmann, and R. Roscher. Automatic differentiation of damaged and unharmed grapes using rgb images and convolutional neural networks. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 347–359. Springer, 2020.
- [15] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the opportunities and risks of foundation models, 2022.
- [16] G. Bradski. The OpenCV Library. *Dr. Dobb’s Journal of Software Tools*, 2000.
- [17] J. F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, November 1986.
- [18] D. P. Cavallo, M. Cefola, B. Pace, A. F. Logrieco, and G. Attolico. Non-destructive and contactless quality evaluation of table grapes by a computer vision system. *Computers and Electronics in Agriculture*, 156:558–564, 2019.

- [19] H. Chen, W. Li, and Z. Shi. Adversarial instance augmentation for building change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021.
- [20] H. Chen and Z. Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote sensing*, 12(10):1662, 2020.
- [21] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 833–851, Cham, 2018. Springer International Publishing.
- [22] C. Cheng, J. Fu, H. Su, and L. Ren. Recent advancements in agriculture robots: Benefits and challenges. *Machines*, 11(1):48, 2023.
- [23] G. Cheng, J. Han, and X. Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.
- [24] T. Choi, O. Would, A. Salazar-Gomez, and G. Cielniak. Self-supervised representation learning for reliable robotic monitoring of fruit anomalies. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2266–2272. IEEE, 2022.
- [25] T. A. Ciarfuglia, I. Marian Motoi, L. Saraceni, and D. Nardi. Pseudo-label generation for agricultural robotics applications. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1685–1693, 2022.
- [26] T. A. Ciarfuglia, I. M. Motoi, L. Saraceni, M. Fawakherji, A. Sanfeliu, and D. Nardi. Weakly and semi-supervised detection, segmentation and tracking of table grapes with limited and noisy data. *Computers and Electronics in Agriculture*, 205:107624, 2023.
- [27] T. A. Ciarfuglia, I. M. Motoi, L. Saraceni, and D. Nardi. Can robots “taste” grapes? estimating ssc with simple rgb sensors. *Computers and Electronics in Agriculture*, 239:110845, 2025.
- [28] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.
- [29] L. Coviello, M. Cristoforetti, G. Jurman, and C. Furlanello. Gbcnet: In-field grape berries counting for yield estimation by dilated cnns. *Applied Sciences*, 10(14), 2020.
- [30] M. Dalmaso, F. Garcia-Ruiz, T. A. Ciarfuglia, L. Saraceni, I. M. Motoi, E. Gil, and A. Sanfeliu. Human-robot harvesting plan negotiation: Perception,

- grape mapping and shared planning. In *2024 7th Iberian Robotics Conference (ROBOT)*, pages 1–8, 2024.
- [31] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2115–2118. Ieee, 2018.
 - [32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
 - [33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.
 - [34] L. Droukas, Z. Doulgeri, N. L. Tsakiridis, D. Triantafyllou, I. Kleitsiotis, I. Mariolis, D. Giakoumis, D. Tzovaras, D. Kateris, and D. Bochtis. A survey of robotic harvesting systems and enabling technologies. *Journal of Intelligent & Robotic Systems*, 107(2):21, 2023.
 - [35] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. Sentinel-2: Esa’s optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, 2012. The Sentinel Missions - New Opportunities for Science.
 - [36] N. Dvornik, J. Mairal, and C. Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 364–380, 2018.
 - [37] N. Dvornik, J. Mairal, and C. Schmid. On the importance of visual context for data augmentation in scene understanding. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2014–2028, 2019.
 - [38] D. Dwibedi, I. Misra, and M. Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1310–1319, 2017.
 - [39] H.-S. Fang, J. Sun, R. Wang, M. Gou, Y.-L. Li, and C. Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 682–691, 2019.
 - [40] S. Fang, K. Li, J. Shao, and Z. Li. Snunet-cd: A densely connected siamese network for change detection of vhr images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.
 - [41] M. Fawakherji, C. Potena, I. Prevedello, A. Pretto, D. D. Bloisi, and D. Nardi. Data augmentation using gans for crop/weed segmentation in precision farming.

- In *2020 IEEE Conference on Control Technology and Applications (CCTA)*, pages 279–284, 2020.
- [42] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
- [43] S. Ghaffarian, N. Kerle, and T. Filatova. Remote sensing-based proxies for urban disaster risk management and resilience: A review. *Remote Sensing*, 10(11), 2018.
- [44] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021.
- [45] T. Ginart, M. JinYE Zhang, and J. Zou. Mldemon: deployment monitoring for machine learning systems. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3962–3997. PMLR, 28–30 Mar 2022.
- [46] A. Gongal, S. Amatya, M. Karkee, Q. Zhang, and K. Lewis. Sensors and systems for fruit detection and localization: A review. *Computers and Electronics in Agriculture*, 116:8–19, 2015.
- [47] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, oct 2020.
- [48] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.
- [49] S. Gutiérrez, J. Tardaguila, J. Fernández-Novales, and M. Diago. On-the-go hyperspectral imaging for the in-field estimation of grape berry soluble solids and anthocyanin concentration. *Australian Journal of Grape and Wine Research*, 25(1):127–133, 2019.
- [50] R. Güldenring and L. Nalpantidis. Self-supervised contrastive learning on agricultural images. *Computers and Electronics in Agriculture*, 191:106510, 2021.
- [51] M. Halstead, S. Denman, C. Fookes, and C. McCool. Fruit detection in the wild: The impact of varying conditions and cultivar. In *2020 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, 2020.
- [52] H. Hammouch, S. Mohapatra, M. El-Yacoubi, H. Qin, H. Berbia, P. Mäder, and M. Chikhaoui. Ganset - generating annotated datasets using generative adversarial networks. In *2022 International Conference on Cyber-Physical Social Intelligence (ICCSI)*, pages 615–620, 2022.

- [53] X. Hao, L. Liu, R. Yang, L. Yin, L. Zhang, and X. Li. A review of data augmentation methods of remote sensing image target recognition. *Remote Sensing*, 15(3):827, 2023.
- [54] Z. K. J. Hartley and A. P. French. Domain adaptation of synthetic images for wheat head detection. *Plants*, 10(12), 2021.
- [55] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [56] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [57] N. He, L. Wang, P. Zheng, C. Zhang, and L. Li. Cbsasnet: A siamese network based on channel bias split attention for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024.
- [58] L. Hoyer, D. J. Tan, M. F. Naeem, L. Van Gool, and F. Tombari. Semivl: semi-supervised semantic segmentation with vision-language guidance. In *European Conference on Computer Vision*, pages 257–275. Springer, 2024.
- [59] W. Huang, Y. Shi, Z. Xiong, and X. X. Zhu. Adaptmatch: Adaptive matching for semisupervised binary segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.
- [60] S. Illarionova, S. Nesteruk, D. Shadrin, V. Ignatiev, M. Pukalchik, and I. Osleledets. Object-based augmentation for building semantic segmentation: Ventura and santa rosa case study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1659–1668, 2021.
- [61] H. Iqbal. Harisiqbal88/plotneuralnet v1.0.0, Dec. 2018.
- [62] D. I. Jackson. Factors affecting soluble solids, acid, ph, and color in grapes. *American Journal of Enology and Viticulture*, 37(3):179–183, 1986.
- [63] S. Ji, S. Wei, and M. Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1):574–586, 2018.
- [64] H. Jiang, M. Peng, Y. Zhong, H. Xie, Z. Hao, J. Lin, X. Ma, and X. Hu. A survey on deep learning-based change detection from high-resolution remote sensing images. *Remote Sensing*, 14(7):1552, 2022.
- [65] Y. Jiang, J. Liu, J. Wang, W. Li, Y. Peng, and H. Shan. Development of a dual-arm rapid grape-harvesting robot for horizontal trellis cultivation. *Frontiers in Plant Science*, Volume 13 - 2022, 2022.
- [66] G. Jocher. Ultralytics yolov5, 2020.
- [67] G. Jocher, A. Chaurasia, and J. Qiu. Ultralytics yolov8, 2023.

- [68] E. Kalopesa, K. Karyotis, N. Tziolas, N. Tsakiridis, N. Samarinas, and G. Zalidis. Estimation of sugar content in wine grapes via in situ vnir–swir point spectroscopy using explainable artificial intelligence techniques. *Sensors*, 23(3), 2023.
- [69] A. Kamilaris, A. Kartakoullis, and F. X. Prenafeta-Boldú. A review on the practice of big data analysis in agriculture. *Computers and electronics in agriculture*, 143:23–37, 2017.
- [70] A. Kamilaris and F. X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018.
- [71] C. Karam, M. Awad, Y. Abou Jawdah, N. Ezzeddine, and A. Fardoun. Gan-based semi-automated augmentation online tool for agricultural pest detection: A case study on whiteflies. *Frontiers in Plant Science*, 13, 2022.
- [72] A. Kavvada, G. Metternicht, F. Kerblat, N. Mudau, M. Haldorson, S. Laldaparsad, L. Friedl, A. Held, and E. Chuvieco. Towards delivering on the sustainable development goals using earth observations. *Remote Sensing of Environment*, 247:111930, 2020.
- [73] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [74] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023.
- [75] A. Koirala, K. B. Walsh, Z. Wang, and C. McCarthy. Deep learning – method overview and review of use for fruit detection and yield estimation. *Computers and Electronics in Agriculture*, 162:219–234, 2019.
- [76] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.
- [77] P. Kurtser, O. Ringdahl, N. Rotstein, R. Berenstein, and Y. Edan. In-field grape cluster size assessment for vine yield estimation using a mobile robot and a consumer level rgb-d camera. *IEEE Robotics and Automation Letters*, 5(2):2031–2038, 2020.
- [78] V. Lacotte, T. NGuyen, J. D. Sempere, V. Novales, V. Dufour, R. Moreau, M. T. Pham, K. Rabenorosoa, S. Peignier, F. G. Feugier, R. Gaetani, T. Grenier, B. Masenelli, P. da Silva, A. Heddi, and A. Lelevé. Pesticide-free robotic control of aphids as crop pests. *AgriEngineering*, 4(4):903–921, 2022.

- [79] T. Le Dinh, M.-C. Vu, and G. T. Tran. Artificial intelligence in smes: Enhancing business functions through technologies and applications. *Information*, 16(5), 2025.
- [80] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.
- [81] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [82] D.-H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.
- [83] H. Li, L. Wang, and S. Cheng. Harnu-net: Hierarchical attention residual nested u-net for change detection in remote sensing images. *Sensors*, 22(12):4626, 2022.
- [84] J. Li, D. Chen, X. Qi, Z. Li, Y. Huang, D. Morris, and X. Tan. Label-efficient learning in agriculture: A comprehensive review. *Computers and Electronics in Agriculture*, 215:108412, 2023.
- [85] K. Li, X. Cao, Y. Deng, J. Song, J. Liu, D. Meng, and Z. Wang. Semicd-vl: Visual-language model guidance makes better semi-supervised change detector. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [86] R. Li, R. Wang, J. Zhang, C. Xie, L. Liu, F. Wang, H. Chen, T. Chen, H. Hu, X. Jia, M. Hu, M. Zhou, D. Li, and W. Liu. An effective data augmentation strategy for cnn-based pest localization and recognition in the field. *IEEE Access*, 7:160274–160283, 2019.
- [87] Y. Li, H. Chen, S. Dong, Y. Zhuang, and L. Li. Multi-temporal samplepair generation for building change detection promotion in optical remote sensing domain based on generative adversarial network. *Remote Sensing*, 15(9), 2023.
- [88] M. Lin, Q. Chen, and S. Yan. Network in network, 2014.
- [89] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.
- [90] M. Lippi, M. Santilli, R. F. Carpio, J. Maiolini, E. Garone, V. Cristofori, and A. Gasparri. An autonomous spraying robot architecture for sucker management in large-scale hazelnut orchards. *Journal of Field Robotics*, 41(7):2114–2132, 2024.
- [91] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2):261–318, 2020.

- [92] S. Liu, S. Cossell, J. Tang, G. Dunn, and M. Whitty. A computer vision system for early stage grape yield estimation based on shoot detection. *Computers and Electronics in Agriculture*, 137:88–101, 2017.
- [93] Y. Liu, S. Zhou, Z. Wan, Z. Qiu, L. Zhao, K. Pang, C. Li, and Z. Yin. A self-supervised anomaly detector of fruits based on hyperspectral imaging. *Foods*, 12(14):2669, 2023.
- [94] Y. Lu, D. Chen, E. Olaniyi, and Y. Huang. Generative adversarial networks (gans) for image augmentation in agriculture: A systematic review. *Computers and Electronics in Agriculture*, 200:107208, 2022.
- [95] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152:166–177, 2019.
- [96] S. L. Madsen, M. Dyrmann, R. N. Jørgensen, and H. Karstoft. Generating artificial images of plant seedlings using generative adversarial networks. *Biosystems Engineering*, 187:147–159, 2019.
- [97] M. Miranda, L. Zabawa, A. Kicherer, L. Strothmann, U. Rascher, and R. Roscher. Detection of anomalous grapevine berries using variational autoencoders. *Frontiers in Plant Science*, 13:729097, 2022.
- [98] S. Mittal, M. Tatarchenko, and T. Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1369–1379, 2019.
- [99] I. M. Motoi, V. Belli, A. Carpineto, D. Nardi, and T. A. Ciarfuglia. Synthetic data generation for anomaly detection on table grapes. *Smart Agricultural Technology*, 10:100787, 2025.
- [100] I. M. Motoi, W. Huang, T. A. Ciarfuglia, D. Nardi, and X. X. Zhu. Changemix: Change-aware cut-and-paste augmentation for semi-supervised remote sensing change detection. Manuscript submitted for publication, 2026.
- [101] I. M. Motoi, L. Saraceni, D. Nardi, and T. A. Ciarfuglia. Evaluating the efficacy of cut-and-paste data augmentation in semantic segmentation for satellite imagery. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 9802–9806. IEEE, 2024.
- [102] Nitin, S. B. Gupta, R. Yadav, F. Bovand, and P. K. Tyagi. Developing precision agriculture using data augmentation framework for automatic identification of castor insect pests. *Frontiers in Plant Science*, 14:1101943, 2023.
- [103] K. Noguchi, Y. Takemura, M. Tominaga, and K. Ishii. Mandarin orange anomaly detection with cycle-gan. In *2024 IEEE/SICE International Symposium on System Integration (SII)*, pages 644–649. IEEE, 2024.
- [104] S. Nuske, K. Wilshusen, S. Achar, L. Yoder, S. Narasimhan, and S. Singh. Automated visual yield estimation in vineyards. *Journal of Field Robotics*, 31(5):837–860, 2014.

- [105] L. F. P. Oliveira, A. P. Moreira, and M. F. Silva. Advances in agriculture robotics: A state-of-the-art review and challenges ahead. *Robotics*, 10(2), 2021.
- [106] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [107] R. Padilla, S. L. Netto, and E. A. B. da Silva. A survey on performance metrics for object-detection algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 237–242, 2020.
- [108] F. Palacios, M. P. Diago, P. Melo-Pinto, and J. Tardaguila. Early yield prediction in different grapevine varieties using computer vision and machine learning. *Precision Agriculture*, 24(2):407–435, Apr 2023.
- [109] PaleBlue AS. *VR Farming Environment Specification*. The CANOPIES Project, 2014.
- [110] M. Palumbo, M. Cefola, B. Pace, G. Attolico, and G. Colelli. Computer vision system based on conventional imaging for non-destructively evaluating quality attributes in fresh and packaged fruit and vegetables. *Postharvest Biology and Technology*, 200:112332, 2023.
- [111] C. Park, A. Awadalla, T. Kohno, and S. Patel. Reliable and trustworthy machine learning for health using dataset shift detection. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3043–3056. Curran Associates, Inc., 2021.
- [112] G. Paulin and M. Ivasic-Kos. Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. *Artificial Intelligence Review*, 56(9):9221–9265, 2023.
- [113] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang. Semicdnet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5891–5906, 2020.
- [114] D. Peng, M. Liu, and H. Guan. Perturbation matters: A novel approach for semi-supervised remote sensing imagery change detection. *Remote Sensing*, 17(4):576, 2025.
- [115] D. Peng, Y. Zhang, and H. Guan. End-to-end change detection for high resolution satellite images using improved unet++. *Remote Sensing*, 11(11):1382, 2019.
- [116] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Trans. Graph.*, 22(3):313–318, jul 2003.

- [117] R. Polvara, S. Molina, I. Hroob, A. Papadimitriou, K. Tsiolis, D. Giakoumis, S. Likothanassis, D. Tzovaras, G. Cielniak, and M. Hanheide. Bacchus long-term (blt) data set: Acquisition of the agricultural multimodal blt data set with automated robot deployment. *Journal of Field Robotics*, 41(7):2280–2298, 2024.
- [118] A. Pretto, S. Aravecchia, W. Burgard, N. Chebrolu, C. Dornhege, T. Falck, F. V. Fleckenstein, A. Fontenla, M. Imperoli, R. Khanna, F. Liebisch, P. Lottes, A. Milioto, D. Nardi, S. Nardi, J. Pfeifer, M. Popovic, C. Potena, C. Pradalier, E. Rothacker-Feder, I. Sa, A. Schaefer, R. Siegwart, C. Stachniss, A. Walter, W. Winterhalter, X. Wu, and J. Nieto. Building an aerial-ground robotics system for precision farming: An adaptable solution. *IEEE Robotics Automation Magazine*, 28(3):29–49, 2021.
- [119] R. Pérez-Zavala, M. Torres-Torriti, F. A. Cheein, and G. Troni. A pattern recognition strategy for visual grape bunch detection in vineyards. *Computers and Electronics in Agriculture*, 151:136–149, 2018.
- [120] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.
- [121] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.
- [122] T. Remez, J. Huang, and M. Brown. Learning to segment via cut-and-paste. In *Proceedings of the European conference on computer vision (ECCV)*, pages 37–52, 2018.
- [123] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.
- [124] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.
- [125] R. Roscher, L. Roth, C. Stachniss, and A. Walter. Data-centric digital agriculture: A perspective. *arXiv preprint arXiv:2312.03437*, 2023.
- [126] C. Rother, V. Kolmogorov, and A. Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.
- [127] T. T. Santos, L. L. de Souza, A. A. dos Santos, and S. Avila. Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. *Computers and Electronics in Agriculture*, 170:105247, 2020.

- [128] L. Saraceni, I. M. Motoi, D. Nardi, and T. A. Ciarfuglia. Agrisort: A simple online real-time tracking-by-detection framework for robotics in precision agriculture. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2675–2682. IEEE, 2024.
- [129] L. Saraceni, I. M. Motoi, D. Nardi, and T. A. Ciarfuglia. Self-supervised data generation for precision agriculture: Blending simulated environments with real imagery. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pages 71–77. IEEE, 2024.
- [130] A. Shafique, G. Cao, Z. Khan, M. Asad, and M. Aslam. Deep learning-based change detection in remote sensing images: A review. *Remote Sensing*, 14(4):871, 2022.
- [131] Q. Shu, X. Zhu, L. Wan, S. Zhao, D. Liu, L. Peng, and X. Chen. Cutmix-cd: Advancing semi-supervised change detection via mixed sample consistency. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [132] P. Skrabanek and F. Majerík. Simplified version of white wine grape berries detector based on svm and hog features. In R. Silhavy, R. Senkerik, Z. K. Oplatkova, P. Silhavy, and Z. Prokopova, editors, *Artificial Intelligence Perspectives in Intelligent Systems*, pages 35–45, Cham, 2016. Springer International Publishing.
- [133] L. N. Smith and N. Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019.
- [134] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.
- [135] J. Song, S. Gao, Y. Zhu, and C. Ma. A survey of remote sensing image classification based on cnns. *Big earth data*, 3(3):232–254, 2019.
- [136] S. Stavridis, P. Falco, and Z. Doulgeri. Pick-and-place in dynamic environments with a mobile dual-arm robot equipped with distributed distance sensors. In *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, pages 76–82, 2021.
- [137] L. Strothmann, U. Rascher, and R. Roscher. Detection of anomalous grapevine berries using all-convolutional autoencoders. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 3701–3704. IEEE, 2019.
- [138] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

- [139] The Canopies Consortium. Dimensioning of the real-world scenario for final demo. Available at <https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5e3eb6ac0&appId=PPGMS>, 2021. Accessed: 03/29/2024.
- [140] A. Toker, L. Kondmann, M. Weber, M. Eisenberger, A. Camero, J. Hu, A. P. Hoderlein, Ç. Şenaras, T. Davis, D. Cremers, et al. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21158–21167, 2022.
- [141] N. L. Tsakiridis, N. Samarinas, S. Kokkas, E. Kalopesa, N. V. Tziolas, and G. C. Zalidis. In situ grape ripeness estimation via hyperspectral imaging and deep autoencoders. *Computers and Electronics in Agriculture*, 212:108098, 2023.
- [142] J. Underwood, A. Wendel, B. Schofield, L. McMurray, and R. Kimber. Efficient in-field plant phenomics for row-crops with an autonomous ground vehicle. *Journal of Field Robotics*, 34(6):1061–1083, 2017.
- [143] M. Valerio Giuffrida, H. Scharr, and S. A. Tsaftaris. Arigan: Synthetic arabidopsis plants using generative adversarial network. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2064–2071, 2017.
- [144] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6 2014.
- [145] J. E. van Engelen and H. H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, Feb 2020.
- [146] E. Vrochidou, C. Bazinas, M. Manios, G. A. Papakostas, T. P. Pachidis, and V. G. Kaburlasos. Machine vision for ripeness estimation in viticulture automation. *Horticulturae*, 7(9), 2021.
- [147] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2526, 2019.
- [148] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.
- [149] J.-X. Wang, T. Li, S.-B. Chen, J. Tang, B. Luo, and R. C. Wilson. Reliable contrastive learning for semi-supervised change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

- [150] L. Wang, M. Zhang, X. Gao, and W. Shi. Advances and challenges in deep learning-based change detection for remote sensing images: A review through various learning paradigms. *Remote Sensing*, 16(5):804, 2024.
- [151] Q. Wang, S. Nuske, M. Bergerman, and S. Singh. *Automated Crop Yield Estimation for Apple Orchards*, pages 745–758. Springer International Publishing, Heidelberg, 2013.
- [152] X. Wang and J. Liu. Tomato anomalies detection in greenhouse scenarios based on yolo-dense. *Frontiers in Plant Science*, 12:634103, 2021.
- [153] Z. Wang, J. Zhang, L. Bai, H. Chang, Y. Chen, Y. Zhang, and J. Tao. A deep learning based platform for remote sensing images change detection integrating crowdsourcing and active learning. *Sensors*, 24(5), 2024.
- [154] D. Williams, A. Britten, S. McCallum, H. Jones, M. Aitkenhead, A. Karley, K. Loades, A. Prashar, and J. Graham. A method for automatic segmentation and splitting of hyperspectral images of raspberry plants collected in field conditions. *Plant Methods*, 13(1):74, Nov 2017.
- [155] H. Williams, S. Harvey, B. McGuinness, D. Smith, H. Lim, and M. Duke. The barracuda: A novel cutting mechanism for pruning canes wrapped around wires. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pages 3286–3293, 2024.
- [156] H. Williams, D. Smith, J. Shahabi, T. Gee, M. Nejati, B. McGuinness, K. Black, J. Tobias, R. Jangali, H. Lim, M. Duke, O. Bachelor, J. McCulloch, R. Green, M. O'Connor, S. Gounder, A. Ndaka, K. Burch, J. Fourie, J. Hsiao, A. Werner, R. Agnew, R. Oliver, and B. A. MacDonald. Modelling wine grapevines for autonomous robotic cane pruning. *Biosystems Engineering*, 235:31–49, 2023.
- [157] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
- [158] Z. Xia, D. Wu, P. Nie, and Y. He. Non-invasive measurement of soluble solid content and ph in kyoho grapes using a computer vision technique. *Anal. Methods*, 8:3242–3248, 2016.
- [159] Y. Xiong, Y. Ge, L. Grimstad, and P. J. From. An autonomous strawberry-harvesting robot: Design, development, integration, and field evaluation. *Journal of Field Robotics*, 37(2):202–224, 2020.
- [160] Y. Xiong, C. Peng, L. Grimstad, P. J. From, and V. Isler. Development and field evaluation of a strawberry harvesting robot with a cable-driven gripper. *Computers and Electronics in Agriculture*, 157:392–402, 2019.
- [161] M. Xu, S. Yoon, A. Fuentes, and D. S. Park. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, page 109347, 2023.

- [162] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7236–7246, 2023.
- [163] L. Yang, Z. Zhao, and H. Zhao. Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [164] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.
- [165] L. Zabawa, A. Kicherer, L. Klingbeil, A. Milioto, R. Topfer, H. Kuhlmann, and R. Roscher. Detection of single grapevine berries in images using fully convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [166] L. Zabawa, A. Kicherer, L. Klingbeil, R. Töpfer, H. Kuhlmann, and R. Roscher. Counting of grapevine berries in images via semantic segmentation using convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 164:73–83, 2020.
- [167] Z. Zhai, J. F. Martínez, V. Beltran, and N. L. Martínez. Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, 170:105256, 2020.
- [168] C. Zhang, L. Wang, S. Cheng, and Y. Li. Swinsunet: Pure transformer network for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.
- [169] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.
- [170] P. Zhang, D. Yan, X. Cai, Y. Chen, L. Luo, Y. Pan, and X. Zou. Multidirectional dynamic response and swing shedding of grapes: An experimental and simulation investigation under vibration excitation. *Agronomy*, 13(3), 2023.
- [171] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.
- [172] Z.-H. Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 08 2017.
- [173] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

- [174] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.