



Delft University of Technology

## **“From iMage to Market”: Machine-Learning-Empowered Fruit Supply**

Wen, J.

### **DOI**

[10.4233/uuid:590d36fd-8f63-48fe-abea-c0b559173cfa](https://doi.org/10.4233/uuid:590d36fd-8f63-48fe-abea-c0b559173cfa)

### **Publication date**

2025

### **Document Version**

Final published version

### **Citation (APA)**

Wen, J. (2025). *“From iMage to Market”: Machine-Learning-Empowered Fruit Supply*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:590d36fd-8f63-48fe-abea-c0b559173cfa>

### **Important note**

To cite this publication, please use the final published version (if applicable).  
Please check the document version above.

### **Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

### **Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights.  
We will remove access to the work immediately and investigate your claim.

# "From iMage to Market": Machine-Learning-Empowered Fruit Supply

Junhan Wen





# **“FROM IMAGE TO MARKET”**

**Machine-Learning-Empowered Fruit Supply**





“FROM IMAGE TO MARKET”  
**Machine-Learning-Empowered Fruit Supply**

**Proefschrift**

ter verkrijging van de graad van doctor  
aan de Technische Universiteit Delft,  
op gezag van de Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,  
voorzitter van het College voor Promoties,  
in het openbaar te verdedigen op  
donderdag 30 oktober 2025 om 15:00 uur

door

**Junhan WEN**

Master of Science in Integrated Building Systems,  
Eidgenössische Technische Hochschule Zürich, Zwitserland,  
geboren te Wuhan, China.

Dit proefschrift is goedgekeurd door de promotoren.

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof.dr. M.M. de Weerd, T.	Technische Universiteit Delft, promotor
Dr. T. E. P. M. F. Abeel,	Technische Universiteit Delft, promotor

*Onafhankelijke leden:*

Dr. J.C. van Gemert,	Technische Universiteit Delft
Dr. T. Guns,	Katholieke Universiteit Leuven, België
Prof.dr.ir. G.J.P.M. Houben,	Technische Universiteit Delft
Prof.dr.ir. T. Keviczky,	Technische Universiteit Delft
Dr. G.W. Kootstra,	Wageningen University & Research

Dit project ontvangt financiële steun van de Topsector Tuinbouw & Uitgangsmaterialen. Binnen de Topsector werken bedrijfsleven, kennisinstellingen en de overheid samen aan innovaties op het gebied van duurzame productie van gezond en veilig voedsel en de ontwikkeling van een gezonde, groene leefomgeving. We erkennen ook de financiering van Stichting Hagelunie en Interpolis.

SIKS Dissertation Series No. 2025-58

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.



*Keywords:* Machine Learning, Computer Vision, Decision Modeling, Optimization, Phenotypic Growth Modeling, Agri-Food Supply Chain Optimization

*Printed by:* Proefschriftspecialist

*Front & Back:* Designed by Wenpei Li

Copyright © 2025 by Junhan Wen

ISBN 978-94-6518-159-2

An electronic version of this dissertation is available at <http://repository.tudelft.nl/>.

*To my 2021-2025:*

“时光会把你雕刻成你应有的样子。”

/“Time will shape you into who you are meant to be.”



# Contents

<b>Summary</b>	<b>xi</b>
<b>Samenvatting</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Research Objectives. . . . .	2
1.2.1 Fruit scale . . . . .	3
1.2.2 Greenhouse scale . . . . .	6
1.2.3 Market scale. . . . .	7
1.2.4 Shared prerequisites. . . . .	9
1.3 Outline of the Thesis . . . . .	9
<b>2 Datasets Overview</b>	<b>11</b>
2.1 Design of Data Flow. . . . .	12
2.2 Data Collection . . . . .	13
2.3 Dataset Statistics. . . . .	17
<b>3 Long-Term Tracking of Multiple Objects with Biological Development</b>	<b>23</b>
3.1 Introduction. . . . .	24
3.2 Related Work . . . . .	26
3.2.1 Image datasets for multiple-object tracking . . . . .	26
3.2.2 Image datasets for plant science . . . . .	26
3.2.3 Algorithms for multiple-object tracking . . . . .	27
3.3 The <i>Growing Strawberries</i> Dataset ( <i>GSD</i> ) . . . . .	27
3.3.1 Data collection setup . . . . .	28
3.3.2 Ground-truth annotation . . . . .	28
3.3.3 Data characterization . . . . .	30
3.4 Benchmark Studies . . . . .	31
3.4.1 Application of MOT algorithms on <i>GSD</i> . . . . .	31
3.4.2 Assessing comprehensive MOT performance. . . . .	32
3.4.3 Detection performance and impact . . . . .	32
3.4.4 Association performance and impact . . . . .	34
3.4.5 Evaluating results from a downstream application . . . . .	34
3.5 Conclusion . . . . .	37

<b>4</b>	<b>Predicting Sugariness Using Non-Destructive and Affordable Hardware</b>	<b>39</b>
4.1	Introduction . . . . .	40
4.2	Materials and Methods . . . . .	42
4.2.1	Dataset . . . . .	42
4.2.2	Methodology . . . . .	43
4.3	Experiment Results . . . . .	44
4.3.1	An integrated dataset of growth and harvest quality . .	44
4.3.2	Conceptual experiment design . . . . .	45
4.3.3	Brix prediction models require more than images. . .	48
4.3.4	Models reveal dependencies on environment data. . .	48
4.3.5	Images enable individual prediction with env-data. . .	49
4.3.6	Plant load further facilitate prediction performance. . .	49
4.3.7	Image encoders notably impact model performances. .	50
4.4	Conclusion . . . . .	52
<b>5</b>	<b>Data-Driven Precision Harvest Planning via Visual Growth Forecasting</b>	<b>55</b>
5.1	Introduction . . . . .	56
5.2	Background and Related Work . . . . .	58
5.2.1	Current practice with vision data . . . . .	58
5.2.2	Formulation of growth functions . . . . .	59
5.2.3	Prediction of future growth . . . . .	59
5.2.4	Harvest and supply optimization . . . . .	59
5.3	Problem Formulation. . . . .	60
5.3.1	Context and Definition . . . . .	60
5.3.2	Formulation. . . . .	60
5.3.3	Data source . . . . .	61
5.4	Methodology . . . . .	61
5.4.1	Processing the images . . . . .	62
5.4.2	Modeling the growths. . . . .	62
5.4.3	Predicting the growths . . . . .	63
5.4.4	Optimizing the harvest plans . . . . .	64
5.5	Experiments and Demonstrations . . . . .	65
5.5.1	Ripeness evaluation . . . . .	65
5.5.2	Growth function formulation . . . . .	66
5.5.3	Growth prediction . . . . .	67
5.5.4	Harvesting optimization . . . . .	69
5.6	Conclusion . . . . .	71
<b>6</b>	<b>Performance and Interaction Assessment of Model Architectures and Bivariate Smart Predict-then-Optimize</b>	<b>73</b>
6.1	Introduction . . . . .	74

6.2	Related Work . . . . .	75
6.2.1	Decision-focused learning . . . . .	75
6.2.2	Optimizations constrained by contracting decisions . . .	76
6.3	Research Questions . . . . .	77
6.4	Methodology . . . . .	77
6.5	Problem and Task Formulation . . . . .	78
6.5.1	Background . . . . .	78
6.5.2	Problem characteristics . . . . .	79
6.5.3	Objective function and specific SPO framework . . . . .	80
6.5.4	Data . . . . .	81
6.6	Experimental Results . . . . .	82
6.6.1	Model architecture impacts performance and training-testing discrepancy on regret . . . . .	82
6.6.2	Models learn alike feature patterns from PFL and DFL . . .	84
6.6.3	Pre-training mitigates discrepancy and reduces regret . . .	85
6.6.4	The sub-problems unveiled as classical DFL tasks . . . .	87
6.6.5	Real-world impact . . . . .	87
6.7	Discussion . . . . .	90
<b>7</b>	<b>Discussion</b>	<b>91</b>
7.1	Answers to Research Questions . . . . .	91
7.2	Outlook and Insights . . . . .	95
<b>A</b>	<b>Appendix</b>	<b>98</b>
A.1	Hosting and licensing information . . . . .	99
A.2	Implementation details for Chapter 2 . . . . .	102
A.3	Implementation details for Chapter 3 . . . . .	103
A.4	Implementation details for Chapter 4 . . . . .	107
A.5	Implementation details for Chapter 6 . . . . .	109
A.6	Further experimental results of Chapter 3 . . . . .	112
A.7	Further experimental results of Chapter 4 . . . . .	116
A.8	Further experimental results of Chapter 6 . . . . .	119
<b>B</b>	<b>Bibliography</b>	<b>127</b>
<b>A</b>	<b>Acknowledgments</b>	<b>151</b>
<b>C</b>	<b>Curriculum Vitæ</b>	<b>155</b>
<b>L</b>	<b>List of Publications</b>	<b>157</b>
<b>T</b>	<b>Titles in the SIKS Dissertation Series since 2016</b>	<b>159</b>





# Summary

Artificial intelligence (AI) has become a widely discussed and transformative technology, with its adoption growing across industries to drive insights and impact. In this thesis, we explore how AI methods and algorithms can facilitate the operation of soft-fruit supply chains, using strawberries as a case study.

The thesis begins by presenting the general background and various perspectives from related works on how AI and machine learning (ML) have been applied to address problems in agricultural or horticultural practices. This includes tasks that, while not directly optimizing supply strategies, still contribute to solving broader challenges. In a nutshell, this thesis categorizes the scope of study into three scales: the single-fruit scale, the greenhouse scale, and the market scale. Within each scale, we review the existing research, identify knowledge gaps, and introduce robust and applicable methodologies capable of dealing with real-world conditions.

Since no publicly available datasets met the requirements of the research plan, we established several datasets for research on the soft-fruit supply chain through collecting, annotating, and (pre-)processing data. These newly curated datasets not only support the research presented in this thesis but also lay a foundation for future research from various perspectives. Details about these datasets are introduced in Chapter 2. Moreover, we conceptualize the process of gathering longitudinal observations from growth monitoring images as a multiple object tracking (MOT) task. We named the image collection and their MOT annotations as “The Growing Strawberries (*GSD*)”. The computer vision challenge that *GSD* brings are further benchmarked and discussed in Chapter 3. Following this, the core contributions of the thesis is presented from Chapter 3 to Chapter 6, each corresponding to a published paper or one currently under review. Finally, Chapter 7 summarizes the research findings, answering the research questions proposed in Chapter 1 and discussing the overall work of the thesis.

We discuss these contributions for each of the three mentioned scales separately:

**At the fruit scale**, we designed and analyzed novel methodologies to keep track of the fruit growths and to predict key properties, including both external characteristics like ripeness and internal qualities such as sweetness. For the ripeness, we propose to use appearance properties, mainly the hue, as an objective metric to quantify it. For the sweetness, we trained deep neural networks to perform non-destructive prediction using environmental and image data, individually and integrally.

Our employment of color analysis and ML models provides a non-destructive and generalizable manner that ensures consistency when upstream and downstream parties in a supply chain estimate the properties of fruits. Meanwhile, the models perform comparatively with laboratory benchmarks even under imperfect, outdoor data collection. We further demonstrated the model in a mobile app to further facilitate adoption in the field.

By benchmarking state-of-the-art MOT algorithms on *GSD*, we illustrated the new challenges that are brought by this use case: first, the MOT objects change appearance

during the tracking due to their biological development, and second, sparse frame rates introduce irregular movements from image to image. We showcased how fruit properties, such as ripeness, change over its life cycle. The results not only provide quantitative measurements that describe the fruit's biological development, but also depict the pain points of current MOT algorithms' predictions. In the meantime, by quantifying these changes over the biological development, we also retrieval relevant information and datasets to support predictions of the changes.

**At the greenhouse scale**, we designed a framework that optimizes the timing of fruit harvesting by integrating the aforementioned quantified changes over biological development, based on sequential demands about the desired quantities to be harvested. Essentially, the framework makes fruit-specific decisions on dates of harvests by leveraging the monitoring data. The decisions are thus made to enhance both current and future demand-fulfillment capabilities. At each stage of this framework, we evaluated various methods and discussed their effectiveness in achieving the stage targets. For example, how to process the infield data to achieve coherent functions about the ripeness development, how to predict future changes, how to include different perspectives in the optimization model, and etc. As the decisions are made for each specific fruit, the work also demonstrates significant potential for integration with mobile apps and harvesting robots. On top of that, the information retrieval function can also serve as a standalone application to provide objective fruit-level quality assessment.

**At the market scale**, we focus on the portfolio optimization of a grower under a widely applied mechanism of the market system: the majority of demands for harvests are predetermined through advance contracts, which also serves as an a priori condition of the solution proposed at the greenhouse level. The local market, with dynamic prices and demands, can be used to save losses from the difference in contracted demands and the actual yield. To mitigate outlying decision failures, we introduced the "smart predict-then-optimize (SPO)" method, which trains models to predict future yield and local market prices. Our results illustrate that SPO loss primarily affects the bias layer in neural networks, contrasting with models trained using mean squared error (MSE). This difference essentially leads to more conservative estimations in decision-making scenarios, and also motivates and highlights the importance of effective MSE-based pre-training. Additionally, our study reveals how SPO loss makes models interact when multiple neural networks are trained to predict decision parameters with diverse functions. This insight expands the applicability of SPO loss across a broader range of use cases and model architectures, underscoring its contribution to the field of decision-focused learning.

In conclusion, this thesis introduces diverse data-driven methodologies to tackle the distinct tasks involved in optimizing fruit supply, using strawberries as a case study. Central to our approach is the effective utilization of data, which serves as the foundation for solutions that span from fruit-level evaluations to market-level planning. By leveraging analytics of non-destructive data, our solutions provide objective estimations of fruit quality, fostering a more consistent shared understanding between sellers and buyers while reducing potential food waste. Overall, these advancements push the boundaries of AI in supporting decision-making during the supply of soft fruits, particularly for smaller growers. The findings not only empower more efficient and sustainable supply chain operations but also highlight the strong potential for many practical real-world applications.

# Samenvatting

**A**rtificial intelligence (AI) is uitgegroeid tot een veelbesproken en transformatieve technologie, waarbij de toepassing ervan in steeds meer sectoren inzichten en impact oplevert. In dit proefschrift onderzoeken we hoe AI-methoden en algoritmen de werking van productieketens voor zacht fruit kunnen vergemakkelijken, met aardbeien als case study.

De dissertatie begint met het presenteren van de algemene achtergrond en verschillende perspectieven van gerelateerde werken over hoe AI en machine learning (ML) zijn toegepast om problemen in de land- en tuinbouw aan te pakken. Dit omvat taken die, hoewel ze niet direct voorzieningsstrategieën optimaliseren, toch bijdragen aan het oplossen van bredere uitdagingen. In een notendop categoriseert dit proefschrift het studiegebied in drie schalen: de schaal van één vrucht, de schaal van de kas en de schaal van de markt. Binnen elke schaal beoordelen we het bestaande onderzoek, identificeren we open vragen en introduceren we robuuste en toepasbare methodologieën die kunnen omgaan met de omstandigheden in de echte wereld.

Omdat er geen publiek beschikbare datasets waren die voldeden aan de eisen van het onderzoeksplan, hebben we verschillende datasets opgesteld voor onderzoek naar de productieketen van zacht fruit door data te verzamelen, te annoteren en te bewerken. Deze nieuwe datasets ondersteunen niet alleen het onderzoek dat in dit proefschrift, maar leggen ook een basis voor toekomstig onderzoek vanuit verschillende perspectieven. Details over deze datasets worden geïntroduceerd in Chapter 2. Bovendien conceptualiseren we het proces van het verzamelen van longitudinale observaties van groeimonitoringbeelden als een meervoudige objectvolgtaak (MOT). We noemden de beeldverzameling en hun MOT-annotaties “The Growing Strawberries (GSD)”. De uitdagingen voor de beeldherkeningstaken die dit met zich meebrengt worden verder besproken in Chapter 3. Daarna worden de belangrijkste bijdragen van het proefschrift gepresenteerd van Chapter 3 tot Chapter 6, elk van deze bijdragen correspondeert met een gepubliceerd artikel of een artikel dat momenteel wordt getoetst. Tot slot geeft Chapter 7 een samenvatting van de onderzoeksbevindingen, waarbij de in Section 1.2 voorgestelde onderzoeksvragen worden beantwoord en het algehele werk van het proefschrift wordt besproken.

We bespreken deze bijdragen voor elk van de drie genoemde schalen afzonderlijk:

**Op fruitschaal** hebben we nieuwe methodologieën ontworpen en geanalyseerd om de groei van het fruit bij te houden en belangrijke eigenschappen te voorspellen, waaronder zowel uiterlijke kenmerken zoals rijpheid als inwendige kwaliteiten zoals zoetheid. Voor de rijpheid stellen we voor om uiterlijke kenmerken te gebruiken, voornamelijk de tint, als een objectieve maat om deze te kwantificeren. Voor de zoetheid hebben we neurale netwerken getraind om niet-destructieve voorspellingen te doen met behulp van omgevings- en beeldgegevens, afzonderlijk en integraal.

Ons gebruik van kleuranalyse en ML-modellen biedt een niet-destructieve, generaliseerbare en consistente manier om de eigenschappen van fruit in te schatten voor upstream-

en downstreampartijen in een productieketen. De modellen presteren vergelijkbaar met laboratoriumonderzoeken, zelfs bij imperfecte gegevensverzameling. We hebben het model verder gedemonstreerd in een mobiele app om de toepassing in het veld verder te vergemakkelijken.

Door state-of-the-art algoritmes voor MOT te testen op GSD, illustreerden we de nieuwe uitdagingen die deze toepassing met zich meebrengt: ten eerste veranderen de MOT-objecten (fruit) van uiterlijk tijdens het volgen door hun biologische ontwikkeling, en ten tweede zorgen lage framerate's voor onregelmatige bewegingen van beeld naar beeld. We lieten zien hoe fruiteigenschappen, zoals rijpheid, veranderen gedurende de levenscyclus. De resultaten bieden niet alleen kwantitatieve metingen die de biologische ontwikkeling van het fruit beschrijven, maar tonen ook de pijnpunten van de voorspellingen van de huidige MOT-algoritmen. Door deze veranderingen tijdens de biologische ontwikkeling te kwantificeren, vinden we ondertussen ook relevante informatie en datasets om voorspellingen van de veranderingen te ondersteunen.

**Op kasschaal** hebben we een framework ontworpen dat, gebaseerd op eisen over de gewenste hoeveelheden die geoogst moeten worden, de timing van het oogsten van fruit optimaliseert door de bovengenoemde metingen over de biologische ontwikkeling te gebruiken. Het framework neemt fruitspecifieke beslissingen over oogstdata door gebruik te maken van de monitoringgegevens. De beslissingen worden dus genomen om zowel de huidige als de toekomstige mogelijkheden om aan de vraag te voldoen te verbeteren. In elke fase van dit framework hebben we verschillende methoden geëvalueerd en hun effectiviteit besproken om de doelstellingen van de fase te bereiken. Bijvoorbeeld, hoe verwerk je de veldgegevens om coherente functies over de rijpheidsontwikkeling te verkrijgen, hoe voorspel je toekomstige veranderingen, hoe neem je verschillende perspectieven op in het optimalisatiemodel, enz. Aangezien de beslissingen voor elke specifieke vrucht worden genomen, toont het werk ook een aanzienlijk potentieel voor integratie met mobiele apps en oogstrobots. Bovendien kan de functie voor het ophalen van informatie ook dienen als zelfstandige toepassing voor een objectieve kwaliteitsbeoordeling op fruitniveau.

**Op de schaal van de markt** richten we ons op de optimalisatie van de portfolio van een teler onder een veel toegepast mechanisme van het marktsysteem: de meeste oogstaanvragen worden vooraf bepaald door middel van voorcontracten, wat ook dient als een voorwaarde voor de oplossing die wij voorstellen op kasniveau. De lokale markt, met dynamische prijzen en vraag, kan gebruikt worden om verliezen te besparen die worden veroorzaakt door het verschil in gecontracteerde vraag en de werkelijke opbrengst. Om beslissingsfouten te beperken, introduceerden we de “smart predict-then-optimize (SPO)” methode, die modellen traint om toekomstige opbrengsten en lokale marktprijzen te voorspellen. Onze resultaten laten zien dat SPO-verlies voornamelijk de bias-laag in neurale netwerken beïnvloedt, in tegenstelling tot modellen die zijn getraind om de mean squared error (MSE) te minimaliseren. Dit verschil leidt tot conservatievere schattingen in besluitvormingsscenario's en benadrukt ook het belang van effectieve, op MSE gebaseerde pre-training. Daarnaast onthult onze studie hoe SPO-verliesmodellen op elkaar inwerken wanneer meerdere neurale netwerken worden getraind om beslissingsparameters met verschillende functies te voorspellen. Dit inzicht breidt de toepasbaarheid van SPO uit naar een breder scala aan gebruikssituaties en modelarchitecturen, en onderstreept de bijdrage ervan aan het veld van decision-focused learning.

Concluderend introduceert dit proefschrift verschillende datagestuurde methodologieën om de verschillende taken aan te pakken die komen kijken bij het optimaliseren van het aanbod van fruit, met aardbeienteelt als case study. Centraal in onze aanpak staat het effectieve gebruik van data, die als basis dienen voor oplossingen die variëren van evaluaties op fruitniveau tot planning op marktniveau. Door gebruik te maken van analyses van niet-destructieve data bieden onze oplossingen objectieve schattingen van de kwaliteit van fruit, waardoor een consistentere gedeeld begrip tussen verkopers en kopers ontstaat en potentiële voedselverspilling wordt verminderd. In het algemeen verleggen deze ontwikkelingen de grenzen van AI bij het ondersteunen van de besluitvorming tijdens de levering van zacht fruit, met name voor kleinere telers. De bevindingen maken niet alleen efficiëntere en duurzamere toeleveringsketens mogelijk, maar benadrukken ook het grote potentieel voor veel praktische toepassingen.



## 1

# Introduction

## 1.1 Background

This dissertation presents algorithms, benchmarks, models, and datasets, designated to support decision-making that are relevant for strategizing the supply of soft fruits.

The motivation for this work stems from the significant impact of food waste on land use, climate, and the global economy. Statistics show that approximately 40-50% of root crops, vegetables, and fruits are lost or wasted globally, with losses reaching trillions of dollars [82]. A range of factors contribute to this waste, including inconsistent harvesting criteria, subjective assessments of fruit quality among supply chain stages, and insufficient information shared across the parties in the supply chain. Together, these issues often leads to spoilage before the fruits reach the down-stream buyers or consumers [209]. Furthermore, inaccurate demand forecasts, diverse grading standards, specific customer requirements, and improper storage practices exacerbate post-harvest waste, particularly during periods of high demand [244].

To minimize waste during harvest and related activities, the concept of precision agriculture – also known as “smart farming” – leverages data-driven tools empowered by machine learning (ML) and artificial intelligence (AI) within the agricultural industry [14, 17, 228]. By utilizing sensors, cameras, and drones, growers can monitor critical aspects of crop production such as micro-climates, crop health, maturity status, plant canopies, geographical specialization, etc. The data collected from these technologies enable the development of analytical and predictive tools that assist in automating agricultural practices, optimizing crop and yield management, and enhancing decision-making throughout their participation in the supply chain [125, 147, 226].

For instance, computer vision (CV) techniques provide fundamental support for processing image data, which identifies and monitors the crops. Convolutional neural networks (CNN) can be trained and deployed to analyze observations of crops or their growing environment [30]. This promotes downstream tasks such as quality inspection [22, 283]. The model can also serve as a part of the automatic control systems, which are increasingly incorporated in agricultural machinery to enhance cost and labor efficiency [245]. Or, more often, detection models such as the region-based CNN (R-CNN) family and the YOLO family are widely used in these type of and branched tasks, i.e. detecting, distin-



guishing, and locating fruits, etc [215, 258, 261]. The outputs of the model can also stream into further crop management [147, 300] and yield estimation [22, 175] tasks.

At a higher level, in terms of optimizing marketing strategies and supply-chain positioning, the employment of mathematical models and operations research methods is widely explored [7, 182, 236]. Before and during harvest, growers must make critical decisions regarding planting and harvesting plans [63]. When participating in a supply chain, they need to estimate the maximum quantities available for sale and decide the minimum contractual quantities by considering the price of each variety [125]. The implementation of such decision-support systems has enabled companies to plan their production more consistently and professionally, leading to enhanced operational efficiency [43].

Most existing research approaches the problem from a single scale, either focusing on point measurement data to optimize overall production and demand strategies or using laboratory data to model the quality of individual crops. However, the real-world practice of optimizing supply strategies and implementing them effectively requires a more comprehensive, interdisciplinary approach. This involves an interconnected, bottom-up flow of data across scales. This gap in the current methodologies is a key barrier preventing state-of-the-art techniques from being successfully applied in practice.

Building on the success of AI and ML algorithms, the vision of the research is to leverage data-driven tools to support the decision-making and optimize the operations at various stages of fruit supply. These tasks include monitoring the cultivation in field, assessing the quality of harvests, predicting the future development of crops, forecasting the demands from downstream parties in the supply chain, and strategizing the supply by assigning the agricultural products to different markets.

In this research, we focus on soft-skinned fruits, also commonly named *soft fruits*, as the main instance. Soft fruits like strawberries and raspberries are popular commercial crops, but are particularly vulnerable due to limited surface protection [4, 293]. Subsequently, these fruits must be harvested at optimal ripeness to ensure that the quality can be guaranteed after post-harvest processes such as transportation and storage, which vary for different markets. Addressing these challenges necessitates optimizing the decisions of their supply. As a typical kind of soft fruit with large consumption, strong flavor, rich nutritional value, and a long harvesting season over the year [244, 250, 281], we chose strawberry as a representative example of the studies in the thesis.

As such, the primary goal of the PhD research is to understand where and how well AI can facilitate strawberry supply from a grower's perspective by investigating and developing algorithms and models to utilize available data from greenhouses' daily operations. The motivations of the research questions comes from concerns directly related to routine practices of growers: *What is the quality of a strawberry? How will this change over time? When to harvest the fruits? How large is a future harvest, or how large shall it be? To which market shall we offer the harvests? And additionally, what supporting data are necessary?*

## 1.2 Research Objectives

The primary objective of this research is to develop data-driven methods for optimizing fruit harvesting decisions in greenhouses, thereby better aligning the supply with downstream. From a practical standpoint, our primary focus is on in-field data collected via reliable and non-intrusive devices, such as observations of fruit growth and micro-climate

sensor readings. In addition, the research is contextualized and supported by external information, including market records and insights into specific supply chain mechanisms.

Considering the abovementioned concerns from practices, we have established eight research questions to identify key tasks for employing data-driven techniques and evaluate their advantages over existing methods. In this way, we addressed the broader issue of supply optimization across three interconnected levels: the fruit scale, the greenhouse scale, and the market scale. In specific, at the fruit scale, we developed deep learning models to analyze, estimate, and predicted the quality and specific attributes of individual fruits. At the market scale, we investigated how factors such as demand, purchasing power of buyers, specific policies, and modeling methods affect the supply and selling strategy. Just like assembling a jigsaw puzzle from its edges to the center, we integrated information from both fruit and market levels to tackle tasks at the greenhouse scale. In specific, we guided harvest decisions by aligning of fruit development insights with downstream demands.

This section introduces and motivates the research questions that arise when we approach the sub-tasks at these three scales in specific, and what the supporting datasets that we have been establishing.

## 1.2.1 Fruit scale

### “What is the quality of a strawberry?”

Traditionally, horticulturalists use the appearance, mainly the skin color, to identify the ripeness level of a fruit [200, 227, 294]. Assessing it subjectively can lead to inconsistent quality interpretations between suppliers and downstream customers [96, 299]. Further, as doing this for all the crops can be time-inefficient and labor-intensive, it also limits the potential to plan harvests and post-harvest activities in time [274]. Following such concerns, researchers trained ML models to estimate the ripeness of fruits, by which a few datasets were also established [76, 287]. This can be combined with detection models to directly distinguish the fruit into different ripeness categories from images, and thus scaling up the evaluation to the whole crops becomes a relatively easier task [242, 293].

In addition to ripeness, fruit characteristics such as size, shape, firmness, sweetness, and shelf life are critical indicators of fruits' quality. Among these, firmness and sweetness are often measured using destructive tests [1, 177]. To overcome the limitations of these methods, many studies have explored ML techniques for non-destructive estimation of these characteristics.

However, most research is conducted under controlled laboratory conditions that involve sealed apparatuses, uniform lighting, and specialized sensors like hyper-spectral imagery (HSI) to assess various fruit attributes [153, 171, 176, 293]. While effective, the reliance on such controlled laboratory setups necessitates a “harvest-first, analyze-second” approach. This limits the practical applicability of these methods for crucial in-field tasks like yield estimation or harvest planning. This thesis therefore pivots to an “analyze-first” approach, focusing on methods compatible with affordable, non-destructive, in-field data collection. The viability of moving beyond the lab is supported by evidence that the most decisive features for quality prediction lie within the visible light spectrum, making expensive specialized sensors not a hard requirement for all time [86, 102, 274, 293]. This fundamental challenge of practical in-field analysis motivates the first research question:

I. How accurately can key quality attributes of fruit, such as external characteristics like ripeness and internal properties like sugariness, be estimated using in-field data? How do these estimations compare to traditional assessments or benchmark methods based on laboratory data?

To address this need, we developed deep learning models for predicting fruit quality at the individual level. These models estimate internal properties non-destructively, such as sugar content (sugariness), which is typically measured by °Brix (or shortened as Brix) in the juice of the fruit. By leveraging in-field data, our approach follows the “analyze-first, harvest-second” principle, which enables continuous quality assessment throughout the entire growth cycle and makes our models ideal for integration into larger agricultural solutions. This practicality is enhanced by their ability to operate offline, allowing for them to be embedded in mobile apps for real-time quality estimation. A small demonstration of such an app was presented at BNAIC 2023 [265]. In addition to Brix, we successfully applied a similar methodology to predict firmness, as discussed in a paper by Jol, Wen, and Van Gemert [123].

Meanwhile, while image data remains essential for accurate fruit-level predictions, the fruit quality is heavily influenced by the growing conditions such as the air temperature, CO<sub>2</sub> level, precipitation, etc. [72, 127, 187, 264]. This means that predictive models often rely on a mix of inputs, from direct sensor measurements to external forecasts. To our knowledge, no prior work has systematically analyzed how fusing data from these different domains, such as combining real observations with predictive forecasts, impacts model performance. This unexamined domain motivates our second research question:

II. How important are the image and micro-climate sensor data in training proper (internal) quality prediction models? What is an optimal way of utilizing this multi-modal data?

In response to this question, we conducted an in-depth evaluation of the practicability of using a single type of data versus integrating multiple data sources for predicting strawberries' Brix. Above all, we introduced a conceptual framework designed to train and validate various approaches to data fusion in this context. Our methodology emphasizes the selection of the optimal model-training process based on both prediction accuracy and practicability. The main findings from our study are detailed in a journal paper by Wen, Abeel, and de Weerd [264] (Chapter 4). We also explored other potential data sources, such as the near-infrared data, which are discussed in the work of Jol, Wen, and Van Gemert [123].

### **“How will this change over time?”**

For strawberry growers, harvesting at the precise peak of quality is crucial for market success. Achieving this requires carefully tracking how key attributes like sweetness and firmness develop on the plant. A common research method to know how the quality develops over-time involves performing destructive tests on batches of strawberries at specific

stages of growth, typically determined by the number of days since flowering [124, 178]. These tests provide a general estimation of the fruit’s development, however, just as “no two snowflakes are alike”, they cannot capture the unique life-cycle of an individual strawberry. Relying on such generalized patterns for decision-making overlooks annual variations in growing conditions, which limits the practical applicability of any optimization models based on them [184].

Given the utility of image data for non-destructive quality estimation, a logical extension is to continuous monitor to enable lifelong and/or real-time assessments of each fruit. Data collected in this way is crucial for supporting yield prediction, harvest planning, and supply management [11, 37, 115, 203]. However, achieving this level of precision agriculture – making decisions for each specific fruit – requires tracking individuals over time [147]. While multiple-object tracking (MOT) is the primary technology for such tasks, its application to plant growth presents a novel problem. Originally developed for tracking dynamic objects like pedestrians, MOT has been extended to domains like live-stock monitoring, as the core task still revolves around tracking motion in real-time over short durations [91, 285]. In contrast, long-term growth monitoring involves static objects where the key difficulty shifts from tracking movement to accurately identifying individuals despite gradual but significant changes in appearance. The performance of conventional MOT algorithms on this distinct type of task is a largely unexplored area, compounded by the lack of relevant public datasets. This fundamental difference between the two domains motivates our next research question:

III. What are the distinct characteristics of a long-term monitoring task compared to typical MOT tasks? How will they affect the performance of existing MOT algorithms, compared to their benchmarks on typical MOT tasks?

In response to this question, and to address the lack of relevant public datasets, our first contribution is the creation and publication of a new benchmark dataset, “The Growing Strawberries” (*GSD*). Using this dataset, we conduct a comprehensive comparison against a typical MOT dataset, *MOT20*, analyzing differences in object characteristics, inter-object relationships, and trajectory properties. This comparison highlights two major challenges that are unique to long-term growth monitoring: significant appearance changes in objects over time and irregular, non-motion-based trajectories. Finally, we benchmarked a variety of MOT algorithms on *GSD* to quantitatively assess their performance on this novel tracking task.

Essentially, the principal objective of applying MOT to growth monitoring data is to understand how fruit quality develops over time. This involves an essential additional step: mapping the series of longitudinal observations to stages of realistic biological developments. However, many existing studies often simply used discrete ripeness categories, such as “flower”, “immature”, “young”, and “mature”. While such labels enable multi-class classification beyond simple “fruit” vs. “non-fruit” detection [115, 242], their discrete nature inherently limits models to basic classification tasks [18, 153, 177, 186, 281] preventing both the progressive assessment of growth and the modeling of continuous biological development. This limitation has practical consequences; for instance, misclassifying an

occluded fruit based on a simple category can cause errors, which can still cascade through long-term forecasts [54, 198]. The need to explore a more refined, preferably continuous, maturity labeling system to mitigate such errors and on downstream tasks raises our next research question:

iv. Apart from MOT metrics, how can monitoring results be practically assessed? In what ways can they be combined with in-field quality analysis to effectively represent biological development?

To answer this, we revisited the traditional criteria for assessing ripeness, which relies on a subjective perception of the fruit's appearance. To create a more objective and quantitative alternative, we proposed using hue as a measurable indicator to track color change. Specifically for strawberries, we employ the A\* channel from the CIELAB color space, which is along the green-to-red color axis, as a straightforward and standard metric for monitoring ripeness development [221]. By mapping this quantitative metric to longitudinal observations, we can track the growth pattern of each individual fruit and compare these patterns across different seasons and treatments. We validated this metric against traditional ripeness labels from open-source datasets to analyze its advantages. Furthermore, this detailed mapping enables a precise analysis of how appearance changes in GSD impact the performance of MOT algorithms.

## 1.2.2 Greenhouse scale

### “When to harvest the fruits?”

Strawberries are non-climacteric fruits, meaning that once harvested, they stop ripening, which halts the development of key quality attributes like appearance, sweetness, and firmness [93, 178]. This standard practice, however, often creates a costly misalignment with fluctuating downstream demand [63, 126]. Optimizing the harvest schedule therefore presents a crucial opportunity to not only increase profitability by preventing economic losses but also to reduce food waste, creating a compelling case for the development of algorithmic decision-support tools [160, 231].

Most research on supply chain harvesting decisions treats quality as an aggregate value. For instance, models might use generalized ripeness states or analyze seasonal patterns to maintain a consistent average quality, as seen in optimizations for orange juice production [184]. However, this way of modeling leaves a gap in practical application: it cannot guide real-world decisions about “which fruit to harvest and which not” because it lacks actual growth tracks of individual fruits. With the advancement of CV techniques, monitoring each fruit's unique development is now possible, opening the door to precision agriculture where decisions can be made for every single fruit [147]. This raises an interesting question:

v. How can we analyze and apply individual growth patterns at the fruit level to support the decisions and optimization at the level of a whole greenhouse?

To answer this, our method builds a predictive pipeline founded on the objective indicator developed in our answer to Research Question IV, i.e. hue or  $A^*$ , to longitudinal observations of each fruit. This creates quantified growth records that detail the development of individual strawberries over time. This process creates quantified growth records for each fruit, detailing their individual development over time. From these historical records, we develop algorithms to predict the future growth trends of newly developing fruits. To enhance the accuracy of this entire pipeline, we employ advanced image processing techniques. Subsequently, these individual growth forecasts are used by optimization algorithms to determine harvest plans, maximizing yield and profitability.

### **“How large shall a future harvest be?”**

Accurate yield prediction is crucial for optimizing both resource management and supply chain positioning [141, 160, 236]. Current yield prediction methods mostly fall into two main categories: forecasting with micro-climate data [232, 248] and estimating based on fruit counting [98, 101, 192]. The former approach uses environmental data to forecast aggregate yields (e.g., in tons) for market-level decisions, such as supply and financial planning [7]. The latter, which increasingly automated by computer vision, provides detailed estimates of fruit numbers for tactical harvest decisions [19, 180, 190, 198]. While both approaches primarily useful insights within their respective scopes, there remains a gap in linking these predictions to actionable decisions. Creating a dynamic mechanism to translate strategic plans into tactical outcomes therefore represents a key opportunity to improve decision-making, raising an interesting research question:

VI. How can individual fruit growth models be used to plan the harvest and optimize the greenhouse’s production?

To answer this, we propose a two-stage framework that connects long-term strategic planning with short-term operational reality. The first stage addresses the strategic definition of supply targets, which are often fixed months in advance due to commercial practices like contract farming. The second, operational stage then focuses on using individual fruit growth models to create precise harvest plans that meet these overarching goals. Our approach aims to synthesize these two perspectives into a single, robust, and practical optimization strategy.

## **1.2.3 Market scale**

### **“To which market shall we offer the harvests?”**

To reduce economic loss and food waste, the soft-fruit industry is adopting more sustainable and economically viable practices in cultivation, harvesting, and supply chain operations [7, 31, 37, 209]. From a grower’s perspective, a key component of this is optimizing their market strategy and operational fulfillment, which involves answering two core questions: first, “how do I invest my potential yield?” and second, “how can I operationally ensure my supply matches that investment?”

The primary financial risk for growers comes from fluctuations in both their production capabilities and market demand [63, 126, 254]. To mitigate market risks, practices

like “contract farming” allow growers to fix production amounts and prices before harvest. However, since production levels are uncertain at the time of contracting, growers typically do not commit their entire expected yield to these contracts, opting instead to keep a portion to sell in local markets [125, 189, 222]. This hedging strategy is further motivated by the asymmetric risk inherent in many supply contracts, where the financial penalties for mismatching the agreed-upon volume are often severe on one side and less consequential on the other, e.g. more costly on under-supplying than over-production, or vice versa. This complex trade-off between the security of contracts and the potential profitability of local or spot markets – a dynamic also seen in industries like energy and finance – highlights the need for context-specific models to strategize supply decisions [194, 236]. This challenge raises a key strategic question:

VII. How can the investment performance be improved by making the yield forecasting model aware of the downstream economic impacts of their predictions? What is the effect on the model and on the profitability?

To account for downstream impacts, we train the predictive model by incorporating the economic consequences of the decisions that result from its forecasts – specifically, the *regret* associated with decision-making. This approach is usually referred to as “decision-focused learning (DFL)”, in contrast to “precision-focused learning (PFL)” of which the primary goal is to maximize forecasting accuracy of prediction models [168]. The key motivation behind using DFL is to effectively manage the different consequences of over- and under-estimation. Since this is a novel application of DFL, we first evaluate its impact on profitability compared to traditional PFL. On top of that, we also investigate how the different loss functions affect the tuning of model hyperparameters.

Once the strategic investment decisions are made, which are informed by yield and market forecasts, the challenge shifts from planning to operational execution. This requires translating these high-level market plans back into specific, fruit-level harvest actions. Using the predictive models and optimization methods developed in our previous work together with growth monitoring data, we can now address this final, capstone research question:

VIII. What is an effective strategy to predict and realize greenhouse production, so as to achieve the best alignment of supply with downstream parties?

To answer this, we propose a two-stage approach that connects long-term strategic planning with short-term operational reality. First, in the strategic stage, we use market data to establish the long-term production targets, specifically, for the supply contracts. Second, in the operational stage, we use the harvest optimization method, supported by our answer to Research Question v, to make precise, short-term adjustments, as the harvest moment approaches. This integration of macro-level planning with micro-level execution ensures the final supply aligns close with long-term targets, or more importantly,



commitments, creating a more robust and actionable production strategy. Consequently, this comprehensive approach answers this final research question by integrating all solutions across the three scales as defined at the beginning of this section.

## 1.2.4 Shared prerequisites

### “What data are desired to support the proposed studies?”

Answering the research questions and developing effective data-driven tools requires comprehensive, multi-modal datasets. Accordingly, our data collection process is designed to combine non-destructive monitoring throughout cultivation with post-harvest analysis of stakeholder-relevant quality attributes.

In practical scenarios, micro-climate information and images are two of the most accessible and non-destructive data types for monitoring the full cultivation process. Research confirms that growth conditions are critical for final fruit quality [98, 146], and images allow for the essential task of tracking individual fruits. This process also generates valuable longitudinal data for advancing research in MOT, where datasets of evolving biological objects are particularly scarce [163, 164, 282, 290].

Our post-harvest data collection encompasses both subjective expert assessments, such as ripeness and marketability, and objective physical measurements. These objective tests include non-destructive metrics like size and shape, as well as destructive analyses of Brix, firmness, and changes during storage. The primary goal is to develop machine learning models that can standardize the inconsistent nature of expert assessments and create a non-destructive alternative to the damaging ground-truth tests. To enable this, we curated a dataset that links the post-harvest quality measurements of each fruit to its corresponding image data from the growth monitoring period. These data create a unique record for each fruit that pairs its visual growth history with its final post-harvest quality attributes.

## 1.3 Outline of the Thesis

To lay the groundwork for this thesis, we begin with a comprehensive exploration of the datasets in Chapter 2. These datasets form the crucial foundation for the proposed solutions and serve as the basis for addressing the research questions outlined in this thesis.

Chapter 3 then focuses on “*The Growing Strawberries*” Dataset (GSD), which is central to growth monitoring. This chapter identifies and validates the unique challenges it introduces to the multi-object tracking (MOT) community. We benchmark state-of-the-art MOT algorithms on GSD, examining how these challenges affect performance and metrics. Additionally, we introduce new methods for evaluating fruit ripeness for the first time in this chapter to demonstrate where current MOT algorithms fall short, emphasizing potential downstream applications. The chapter addresses Research Question III and IV, while also laying the groundwork for potential solutions to Research Question V.

In Chapter 4, we introduce a variety of machine learning methods to predict sugariness, a key internal characteristic of strawberries. Trained on cost-effective, in-field data, our models perform on par with existing benchmarks that use specialized in-lab setups. By estimating post-harvest quality with satisfying accuracy, these models provide growers with consistent and reliable insights to inform their future harvesting decisions. The results provide answers to Research Question I and II.



In Chapter 5, we present an integrated framework to support decision-making for precise strawberry harvests. To align with long-term sales contracts, the framework treats market demand as a fixed quantity, allowing us to scale optimization down to the single-fruit level. Building on the ripeness evaluation and quantification methods introduced in Chapter 3, we developed algorithms and models to forecast ripeness changes over time. These predictions are then aligned with the demand inputs, framing the task as an assignment optimization problem. This chapter advances Research Question IV and VI while addressing Research Question v and viii.

Chapter 6 focuses on the fruit trading system, which shares similarities with the trading of other agricultural products, as well as sectors like renewable energy and stocks. A unique aspect of such a system is that contracts are usually involved in the trading, whilst the production/yield cannot be predetermined at the time of contracting. In this work, we tailor a training method to develop prediction models for yield and market price – both are critical for determining contract amounts. We compare the performance of this method against traditional accuracy-oriented regression models, examining its impact on contracting outcomes and model hyperparameters. Additionally, we propose and validate methods to enhance model performance. Experiments in this chapter utilize open-source data with extensive variables, further validating the method’s applicability beyond strawberry supply. This chapter specifically addresses Research Question VI and VII.

Finally, Chapter 7 summarizes the findings from Chapters 3 to 6 to comprehensively address Research Question I to VIII. We discuss the thesis’s contributions across the three aforementioned research levels – fruit, greenhouse, and market – to systematically support the decision-making scenarios in strawberry supply. We also reflect on the underlying assumptions, unresolved tasks, and suggest potential avenues for future research.

## 2

## 2

## Datasets Overview

For a research project that addresses real-world challenges as reported in this thesis, the *data* serves as a crucial foundation. As such, we collected data from various sources that are integral to the research questions in Section 1.2 and curated a series of datasets by integrating the information from diverse perspectives, ensuring a well-rounded perspective on the problem at hand. Before delving into the main, paper-based chapters as in Chapters 3 to 6, this chapter provides a systematic overview of the datasets that were collected, annotated, and utilized throughout the research.

We begin with the motivation behind establishing these datasets, which is rooted in our overarching methodology to address the aforementioned research questions. Throughout the cultivation process, we collected diverse types of information, including growth-monitoring images, quality measurements, and records of the cultivation environment. In this chapter, we detail the data collection process and provide examples of the datasets, ensuring that the approach is reproducible. The datasets are provided publicly as part of the open science policy, facilitating follow-up research. The chapter also provides a closer look at the datasets themselves, presenting statistical insights and discussing key factors that impact dataset quality. While part of these sections may overlap with Section 3.3 and Section 4.2, which focus on growth tracking and quality prediction, respectively, the purpose here is to highlight the collective significance of these datasets within the broader research framework. To this end, we emphasize not only their individual contributions but also how they interconnect to support the overall goals of the thesis. At the end of this chapter, we present the metadata associated with each dataset, offering practical guidance for potential future users. We aim to facilitate effective use of these datasets in similar or extended research endeavors, helping to advance the field and encourage new solutions to related challenges.

## 2.1 Design of Data Flow

To develop the algorithms and models, the project needs the input of four sources: (i) the images that monitor the cultivation periods, (ii) the environment data in the greenhouse compartment over the same periods, and (iii) destructive test measurements of fruits that are also visible in images, and (iv) data or facts about the soft-fruit market. Figure 2.1 illustrates the data flow by the models involved in this project among the three scopes of studies: fruit level, greenhouse level, and market level. The data were collected during the cultivation periods and after the harvests. Eventually, the data that we have collected or their features or properties after a certain level of processing are fed into the optimization task of harvest planning about specific fruits.

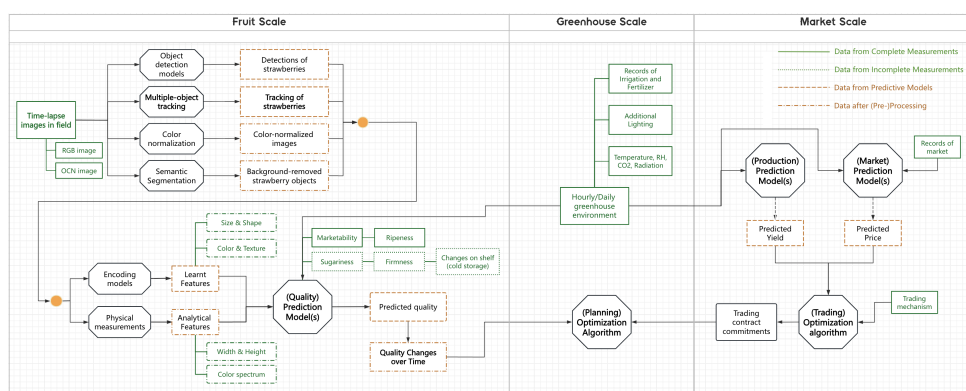


Figure 2.1: Proposed data flow of the models implemented in this research, organized across the three scales of research scope described in Section 1.2. Green boxes represent collected data, while brown boxes indicate data processed during the research. The legend explains the meaning of the box frame line styles. Bolded text highlights the major topics or objects of the paper-based chapters in this thesis.

At the fruit level, we collect two types of data to describe properties of individual fruits: images and measurements. As a non-destructive measuring way, we collect images over the cultivation period, of which the longitudinal observations also form monitoring information about the growth. These images are therefore the main source of growth monitoring and forecasting tasks. Measurements included results of destructive and non-destructive tests, that are labeled or measured by human experiments after the growth. In specific, the sugariness is measured by the percent dissolved solids in aqueous liquids using an Abbe refractometer scaled in Brix units, and the firmness is measured by registering the force required for a Cylinder Probe to penetrate the fruit's flesh to a chosen distance. They are mainly treated as labels for supervised learning using image data.

At the greenhouse level, we collected micro climate information such as temperature, CO<sub>2</sub> and humidity, which is mainly shared over all growths at the same moment. Specifically, for fruit-level quality prediction work, we measured plant load and yield, which was suggested by the greenhouse experts. This information acts as aggregated information for training/anchoring the quality prediction task about individual fruits. Meanwhile, they are also the crucial foundation for the market-level optimization task.

The research at the market level optimizes the matching between gaining from harvests and demands from downstream customers in the supply chain, from the perspectives of a grower. For such, the yield data represents the production quantity and the price and demand over time describe the markets performance. The historical values of all these data are used to train forecasting models, so that the model outputs can be used in planning before the real harvest.

Next, we provide the details on each of these data sources.

## 2.2 Data Collection

### Growth Monitoring Images

According to the research in which hyperspectral images (HSI) techniques are considered, most principal features still fall in to the visual spectrum [85, 102, 274]. Thus, from 2021 to 2023, we used six RGB and OCN<sup>1</sup> cameras (model Survey3N from MAPIR®) to track the growth of 12 *Favori* strawberry plants (each year) in three greenhouses in Netherlands. The cameras were paired in three sets, denoted as *RGB/OCN-1/2/3*. All the images are 4000×3000 pixels. For each pair, the two cameras are installed in parallel with a fixed distance of 10 cm, forming a big view overlap that ensured most fruits were monitored by both cameras. In 2024, we retained the RGB cameras and set up two additional fish-eyed RGB cameras for overhead monitoring, which captured images every three hours. We used this setup to monitor the fruit growths in the same greenhouse as where the data was collected in 2021, so as to form a validation set.

Figure 2.2 gives a sample view of each camera, collected at the same moment. The horizontal shift of view of each camera pair due to the parallel setup can be noticed in each row. Although there exist some small dislocations among the images resulted from the camera shaking from practice, many static reference objects can still be found in the images for re-alignment of the views.

The monitoring of fruit growths is therefore formed by the longitudinal observations, i.e. considering the time-lapse images as frames of a long video, and the fruits are the objects to be tracked over the frames. For instance, the individual bounding boxes (bbox) as shown in each row of Figure 2.3, from which we can form series of observations as in Figure 3.1. For the growth-monitoring images collected in 2021 and 2022, we provide human-annotated bboxes for every strawberry, at all growth stages, along with identity labels for trajectory tracking during the period of their biological developments. This therefore formed a multiple object tracking (MOT) challenge with unique complexities raised by the prolonged monitoring periods. The dataset, published as “*The Growing Strawberries*” Dataset (GSD), is accompanied by a detailed paper, included in this thesis as Chapter 3. The annotation process, i.e. marking the bbox and trajectory IDs as illustrated in Figure 2.2, included an initial annotation stage and three rounds of review and verification from non-annotating workers, so as to mitigate potential labeling errors and personal biases.

On top of the MOT annotation of GSD images, we applied “*Segment Anything*” [133] to segment the strawberries with polygon masks from the bbox images. By connecting the segments from the longitudinal observations from both the RGB and OCN cameras by the stickers in view, we form a dataset of in-field image and post-harvest measurement of

<sup>1</sup>The channels are: Orange/615nm, Cyan/490nm, Near-Infrared/808nm.



Figure 2.2: Example views of the three pairs of cameras. The numbers in the camera name indicate how they were paired. The color-coded rectangles are the ground-truth bbox annotations of object detection. The MOT annotation are noted as the trajectory IDs at the top of the bboxes.

strawberries over three years, as depicted in Figure 2.4. The first-year data is published together with the paper by Wen, Abeel, and de Weerd [264], and the entire collection is available via the *4TU.ResearchData* Platform [271].

As we collected images at different times of the day, we also included exploiting color normalization methods as a crucial step of data pre-processing. The second row in Figure 2.3 gives one example way of normalizing the images – by performing a mixed-use of gamma correction and histogram equalization, inspired by Young-Chang Chang and Reid



[291] and Bezryadin, Bourov, and Ilinih [28]. A few algorithmic color correction methods, intrinsic color measurements, and generative models could be employed for this purpose, of which a further discussion of them can be refer to Chapter 5. In a nutshell, many of the approaches achieved a reduction of the color difference of the time-lapse images, facilitating the computational methods to obtain useful information which quantify the growth of the fruits and can be used in further growth predictions.

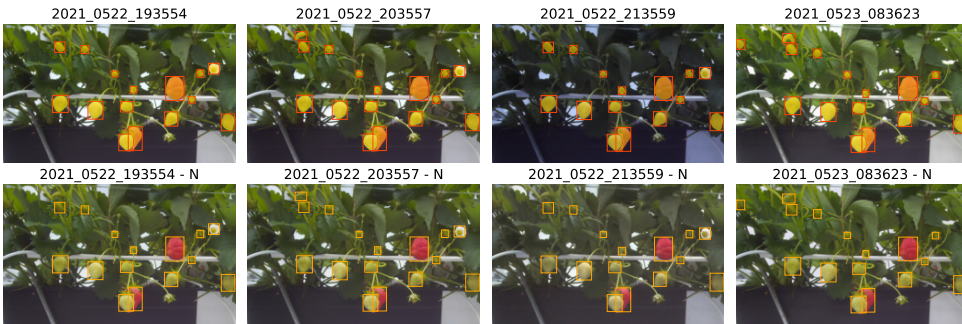


Figure 2.3: Example time-lapse images, along with the detection and segmentation of strawberry fruits. The original names of the images are displayed at the top of the plots, wherein the collection time is indicated. The first row shows the original images, while the second row displays the images after a certain level of color normalization. Detection bounding boxes are shown in both rows, and the segmentations are represented as yellow masks in the first row.

## Quality Attribute Evaluations

A few hours before the harvest, we label a few fruits with an ID by a physical sticker that is visible to the cameras, such as the yellow stickers in Figure 2.2. Then, after it has been harvested, the measurements of the quality attributes can be again connected to the images that were taken in-field, so we formed a dataset to train deep learning models to achieve the goal of “analysis first, harvest second”.

The quality assessment criteria include physical and biochemical indicators of the ripening process, and also the subjective concerns of suppliers of how to categorize the strawberries in the market. The quality attributes are determined through human assessment and device measurements. A horticulturist labels the ripeness of fruit according to the greenhouse standards and decides whether a strawberry is marketable, usually based on its appearance – specifically, its roundness and visible defects if any. Device measurements can be non-destructive or destructive. Strawberries are weighed and measured for size, typically indicated by diameter. Many greenhouses use ring-passing tests to categorize sizes instead of providing exact numbers, with diameter ranges listed in corresponding measurement sheets. Destructive tests assess sugariness (measured in °Brix or written as “Brix” for short), firmness (measured in  $\text{kg}/\text{mm}^2$ ), and changes in appearance and weight loss during a two-week cold storage period. Since sugariness and firmness tests destroy the fruit, the shelf life test is conducted using a different batch from the harvest used for the other measurements.

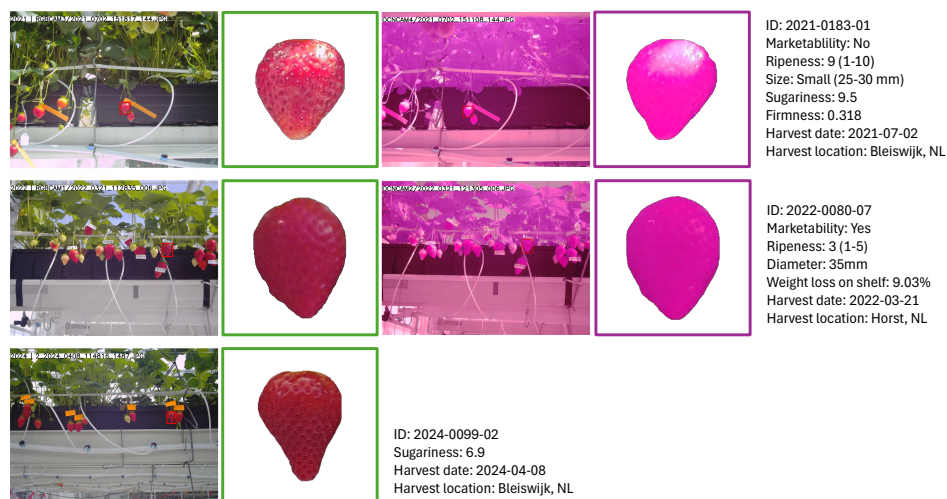


Figure 2.4: Yearly dataset samples: segmented strawberry images labeled with quality properties. The 2021 and 2022 samples contain parallel views from the paired RGB and OCN cameras, offering richer features for analysis.

## Micro-Climate Records

It is demonstrated in many related studies that the greenhouse environment, including the temperatures (at multiple locations), relative humidity (RH), CO<sub>2</sub> density, illumination level, and additional nutrition to the plants influences the growth and final quality significantly [51, 56, 70, 110]. Thus, in our research, the environment information is considered as an auxiliary feature for both prediction models. Sets of sensors monitored various places in the greenhouse or in the neighborhood at a minimum 5-minute frequency.

An overview of the collected attributes are presented by Table 2.1: The sensor measurements of temperatures, RH, CO<sub>2</sub>, and radiation are collected in the greenhouse compartment where the strawberries were cultivated. The rainy condition at the greenhouse's location was recorded as Boolean values, i.e. "whether it rained or not". Additional treatments such as irrigation and cyclic lighting were logged by greenhouse managers. The data were then aggregated into hourly or daily to reduce influences from sensor failures.

## Market Information

To better understand the trading dynamics in the soft-fruit market, we conducted surveys and interviews with several key suppliers among our project stakeholders and reviewed relevant literature. Our qualitative research revealed that most agricultural products, including those we examined, are sold to large distributors through contracts signed weeks or even months in advance. This practice, commonly known as "contract farming", also serves as a risk management tool for growers that helps to mitigate the unpredictability of production. Furthermore, this strategy does not only prevalent in agriculture but is also widely applied across various industries.

Unfortunately, we couldn't access specific data related to strawberries due to confidentiality restrictions. However, when we expanded our scope to encompass other markets,

Table 2.1: Attributes of Growing Environmental Records. The columns indicate the higher-level groups, where “xxx” refers to the name codes of specific (known) locations in the greenhouse. The content of each cell is formatted as “abbreviation: explanation”. The records of indoor micro-climate, outdoor weather, and plat treatment are aggregated hourly in the latest published version of Wen et al. [271].

Abbreviation	Content
INDOOR MICRO-CLIMATE	
PT-xxx	Plant Temperature (PT) at specific locations
LT-xxx	Leaf Temperature (LT) at specific locations
ST-xxx	Substrate Temperature (ST) at specific locations
RT-compt	Room Temperature (GT) in the greenhouse compartment
VT-compt	Ventilation Temperature (VT) in the greenhouse compartment
AH-compt	Absolute Humidity (AH) in the compartment
RH-compt	Relative Humidity (RH) in the compartment
CO2-compt	CO <sub>2</sub> intensity in the compartment
OUTDOOR WEATHER	
Rain	Rained or not
T-outer	Temperature (T) outside
RH-outer	Relative Humidity (RH) outside
AH-outer	Absolute Humidity (AH) outside
RDAN	Net radiation
PAR-xxx	Photosynthetically Active Radiation (PAR) sensor data
PLANT TREATMENT	
Lighting	Artificial lighting by the flowering bulbs
CyLighting	Artificial cyclic lighting by the flowering bulbs
Watering	Amount of irrigation

we found other publicly available online datasets, by Borsa [33] and Jhana [121], that we could still utilized to benchmark the contracting methodologies developed in Chapter 6. In this way, the expanded focus not only enhances the generalizability of our conclusions, but the insights also contribute beyond the strawberry use case.

## 2.3 Dataset Statistics

Based on the collected data and information, we established two collections of datasets: “*Growth monitoring of strawberries*” [270] and “*Quality evaluation of strawberries*” [271]. The growth monitoring datasets consist of time-lapse images used to track the growth of strawberries, along with the corresponding MOT annotations. The quality evaluation datasets include in-field images of strawberries before they were harvested, their associated post-harvest measurements, and relevant environmental information as further inputs for training prediction models.

The annotated growth monitoring images from 2021 and 2022 have been published as “The (original) Growing Strawberries Dataset (GSD)” [269]. We designate the daytime images from 2023 and 2024 as the “*extended-GSD*”, while the remaining unannotated, mostly



dark images are labeled as “*dark-GSD*”. Our exploration on [269] focuses primarily on the properties of its MOT annotations. Moreover, a detailed exploration of *GSD-2021* and a comparison of the characteristics between *GSD-2021* and *MOT20*, a popular dataset from MOTChallenge, is presented in Section 3.3.

For the quality evaluation datasets, we analyzed the label distributions across the different years in which the data was collected. This analysis highlights both the similarities and differences in measurements of various attributes, as well as how these distributions can vary among different cultivation years and locations.

## Growth Monitoring of Strawberries

### Images and annotations

Table 2.2 presents the overall statistics of *GSD* as accessible via [269]. For our benchmarking experiments detailed in Chapter 3, we focused exclusively on the RGB images within *GSD*. Accordingly, we provide the statistics of only these RGB images in Table 3.1 in Section 3.3. On top of that, following three additional rounds review and annotation correction of *GSD*, Table 2.2 has been updated to reflect these revisions, while Table 3.1 remains as it was originally published with the its release in *4TU.ResearchData* [272].

Further, Table 2.3 shows the statistics of the “*extended-GSD*”, which is composed of only growth monitoring images collected in 2023 and 2024. Since both *GSD* and “*extended-GSD*” contain only daytime images, we store the nearly-dark or blacked-out images separately in an entry, designated as “*dark-GSD*”. All three datasets are encapsulated under the collective title, “*Growth Monitoring of Strawberries*”, available on the *4TU.ResearchData* platform.

Table 2.2: Statistical overview of the RGB and OCN images of the latest update of *GSD*. The 2<sup>nd</sup> column lists the duration of data collection. The 3<sup>rd</sup> and 4<sup>th</sup> columns note the amounts of all time-lapse images (“Total images”) and the annotated images that formed the dataset (“Dataset images”), respectively. The last two columns present the total number of annotation bboxes and trajectories.

Camera	Period	Total images	Dataset images	Total bbox	Total tracks
subset: GSD-2021					
<i>RGB-1</i>	Apr 23 - Nov 9	4786	2689	66342	490
<i>RGB-2</i>	Apr 23 - Nov 9	4785	2595	63444	393
<i>RGB-3</i>	Jun 29 - Nov 9	3181	1694	67000	422
<i>OCN-1</i>	Apr 23 - Nov 9	4786	2612	73072	559
<i>OCN-2</i>	Apr 23 - Nov 9	4785	2677	72029	451
<i>OCN-3</i>	Jun 29 - Nov 9	3182	1727	68550	482
subset: GSD-2022					
<i>RGB-1</i>	Feb 22 - Oct 3	5128	3279	93542	562
<i>RGB-2</i>	Feb 22 - Oct 3	4699	3000	115531	895
<i>RGB-3</i>	Feb 22 - Oct 3	5156	3280	108488	753
<i>OCN-1</i>	Feb 22 - Oct 3	5159	3247	72780	690
<i>OCN-2</i>	Feb 22 - Oct 3	5162	3323	100749	766
<i>OCN-3</i>	Feb 22 - Oct 3	5158	3323	107838	749

Table 2.3: Statistical overview of the RGB and OCN images of the *extended-GSD*. The 2<sup>nd</sup> column lists the duration of data collection. The 3<sup>rd</sup> and 4<sup>th</sup> columns note the amounts of all time-lapse images (“Total images”) and the annotated images that formed the dataset (“Dataset images”), respectively.

Camera	Period	Total images	Dataset images
subset: GSD-2023			
<i>RGB-1</i>	May 31 - Aug 29	1819	1413
<i>RGB-2</i>	May 31 - Oct 18	2328	1401
<i>RGB-3</i>	May 31 - Oct 18	2018	1232
<i>OCN-1</i>	May 31 - Aug 29	1592	1109
<i>OCN-2</i>	May 31 - Oct 18	1847	1127
<i>OCN-3</i>	May 31 - Oct 18	1848	1126
subset: GSD-2024			
<i>RGB-1</i>	Feb 6 - Jun 27	3404	2150
<i>RGB-2</i>	Feb 6 - Jun 27	3404	2120
<i>RGB-3</i>	Feb 6 - Jun 27	3404	1998
<i>RGB-4</i>	Feb 6 - Jun 27	3407	2008
<i>Overhead-1</i>	Feb 6 - Jun 27	1136	905
<i>Overhead-1</i>	Feb 6 - Jun 27	1137	906

## MOT trajectories

explain why there is a sample video

Since the cameras are stationary and strawberries generally do not travel long distances throughout their life cycle, many strawberries in *GSD* obtain a complete trajectory from early growth to harvest. For example, Figure 3.1 shows a strawberry positioned on the outer layer, which was largely visible throughout its development. However, not all of the strawberries were completely monitored. As illustrated by the track length statistics in Figure 2.5, a notable portion of tracks are relatively short.

Various factors contribute to these incomplete observations:

1. In cooler weather such as in May or September, the strawberries grow slower and less dense. Both factors make the trajectories naturally longer than in warmer times.
2. The strawberries were growing in dense gathers, as can be observed in the sample views as in Figure 2.2 and the demo video *GSD-sample\_video.mp4* in [269]. For instance, in the branches captured by cameras *RGB-2* and *OCN-2* in Figure 2.2. In these conditions, some strawberries in the inner layers were occluded. Nevertheless, due to varying rates of weight gaining, their positions could shift relative to one another, allowing them to emerge at later stages of maturity in view.
3. The increases in size and weight might also squeeze some strawberries out of frame. For example, the strawberry #397 in the view from *RGB-2* in Figure 2.2; the strawberry #596 in the demo video also moves back and forth at the edge of the frame.

4. Strawberries from the inner layer started to have more complete observations when the outer-layer strawberries were harvested. For example, radical position changes can be observed in the the demo video.
5. Some strawberries grew above the cameras, which were not intended to be monitored. For example, the strawberry #451 in the view from *RGB-2* in Figure 2.2.
6. In addition, human activities, such as harvesting one fruit, can cause nearby fruits to move as well. For example, intentional actions like pulling fruits outward (called “uithalen” in Dutch) may also occur. While these behaviors are not evident in the demo video *GSD-sample\_video.mp4*, they are known to exist.

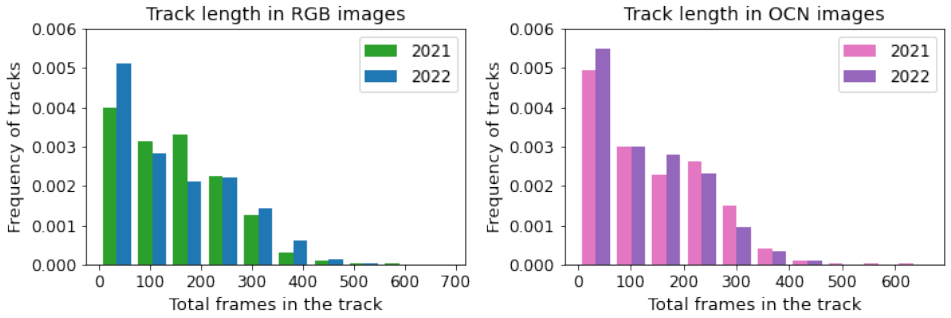


Figure 2.5: Statistics of track lengths in *GSD*, based on the latest update from Wen et al. [270]. The histograms display the statistics for the 2021 and 2022 datasets, with normalized frequencies. Track length is defined as the number of bounding boxes assigned to a track according to the ground-truth MOT annotations.

For instance, in the example subsequence of *GSD* as in *GSD-sample\_video.mp4*, two drastic location changes happened between 12-1 pm, Sep. 1<sup>st</sup>, and between 5-6 pm, Sep. 6<sup>th</sup> because of the harvests. It could be noticed that strawberries #355 and #354 switched positions suddenly between frame 1-2 pm, Sep. 7<sup>th</sup>, because of the harvest of #391. This exemplifies the irregular movements of *GSD* objects. The video is accessible in the *GSD-Sample* file folder in [269].

## Quality Evaluation of Strawberries

### Statistics of measurement data and image labels

Table 2.4 presents the count of data points from measurements conducted between 2021 and 2024. We classify these data points into three different sets due to varying levels of completeness: some data points are excluded from labeling an image because the corresponding strawberry or its physical label in the images were either occluded, poorly labeled, or out of view, leading to gaps in the datasets. Nevertheless, based on the Law of Large Numbers, we expect that a larger amount of quality evaluations will better reveal underlying attribute distributions, which can still be collected regardless of the association with strawberry images. Therefore, we established two dataset configurations: an image-based dataset (noted as “Data points labels an image” as in Table 2.4), where quality

attributes serve as labels for the strawberry segments in the images, and a measurement-based dataset, which archives all measurements associated with a valid sampling date. In addition to that, in selected greenhouses where they also have a regular plan of sampling fruits from harvests, we also collect the measurements for the strawberries of the same species and underwent the same treatment before and during cultivation (hereby, called “same fruit” for short). This approach aims to capture more representative distribution patterns at each harvest round. A time-based comparison of the three sets of data points are as shown in Figure A13.

Table 2.4: Statistical overview of the quality evaluation datasets, organized by the specific attributes. The 3<sup>rd</sup> column presents the total number of fruits that were assessed or measured. The 4<sup>th</sup> column displays the number of data points in the image-based dataset, where we used each quality attribute to label specific strawberry segments in the RGB images. The 5<sup>th</sup> column notes additional measurement data gathered from extra samples used by the greenhouse to assess each harvest.

Quality attribute	Year	# Data points	# Data points labels an image	# Data points from the same treatment
Marketability	2021	315	204	-
	2022	947	758	-
	2023	145	-	-
Ripeness	2021	310	200	-
	2022	947	758	-
	2023	145	-	-
Size	2021	310	199	-
	2022	947	757	-
	2023	145	-	-
Sugariness	2021	1144	199	583
	2022	475	383	475
	2024	1514	269	909
Firmness	2021	245	164	-
Harvest weight	2022	477	373	-
	2023	145	-	-
Weight loss on shelf for 14d	2022	471	373	-

### Distributions of quality evaluations

Figure 2.6 illustrates the label distribution of the images, which are formed by the quality evaluation attributes. We divide the labels into tow rows: three categorical labels in the top row and three continuous labels in the bottom row. As shown, since the harvesting are still based on a market-oriented standard, the majority of strawberry samples are marketable and were harvested at or near optimal ripeness.

Notably, the ripeness distributions differ significantly between 2021 and 2022, which resulted from the distinct labeling criteria used by the two greenhouses. In 2021, one greenhouse rated ripeness on a scale from 1 (green) to 10 (over-ripe), resulting in a pre-dominant score of 7 or 8. In contrast, in 2022, the second greenhouse employed a scale

from 1 to 5, where 1 encompassed all green and greenish strawberries, resulting in most scores clustering around 3 or 4. Meanwhile, as most greenhouses utilized a ring test for measuring size, the highest label categories were therefore limited by the largest ring that they used. Aside from species and cultivation differences, these methodological variations contribute to the differing distribution widths seen in the dataset.

The three continuous attributes in the bottom row are collected from destructive measurements. Sugariness, which we discuss further in Chapter 4, was measured across multiple seasons over three years. According to Figure 2.6, measurements from 2024 are more concentrated to their average, while 2022 measurements show a leftward shift along the x-axis, indicating an overall lower mean value. These distributional differences impact validation performance, highlighting the need for larger datasets to train more generalized models or, alternatively, recommending transfer learning before applying these models to a new greenhouse. Firmness and weight loss (measured after 14 days on the shelf) were assessed in only one year each. For these attributes, our dataset provides a foundational basis for training non-destructive estimation models, with putting the detailed model development and validation for these attributes outside the scope of this thesis.

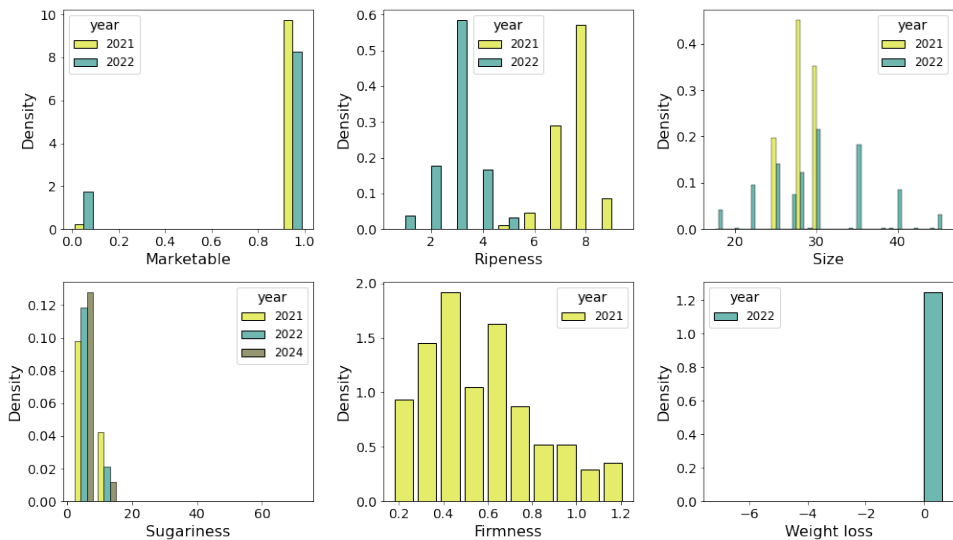


Figure 2.6: Statistics of the measurements (/labels) in the image-based quality evaluation dataset. The quality attribute are noted on the labels of the x-axes respectively. The y values indicate the density of the data that fall within each range over the measurements of the same year.

## 3

## 3

## Long-Term Tracking of Multiple Objects with Biological Development

This chapter explores the unique challenges posed by the “the Growing Strawberries (*GSD*)” to the research of multi-object tracking (MOT). Specifically, it highlights two key challenges: (1) significant appearance changes of objects across frames and (2) irregular object movement patterns. These challenges are illustrated through a comparison with a well-known MOT dataset, *MOT20*, and benchmarking of state-of-the-art MOT algorithms on our dataset. Additionally, we introduce a novel method for quantifying fruit ripeness to highlight the limitations of existing MOT algorithms on *GSD*. While there is some overlap in Section 3.3 with Section 2.2, this chapter emphasizes the new challenges introduced to the MOT field and the reasons behind their emergence.

This chapter was previously published as: Junhan Wen, C. R. Verschoor, C. Feng, I. Epure, T. Abeel, M. de Weerdt. 2024. “The Growing Strawberries Dataset: Tracking Multiple Objects with Biological Development over an Extended Period.” *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2024. [272]

**M**ultiple Object Tracking (MOT) is a rapidly developing research field that targets precise and reliable tracking of objects. Unfortunately, most available MOT datasets typically contain short video clips only, disregarding the indispensable requirement for adequately capturing substantial long-term variations in real-world scenarios. Long-term MOT poses unique challenges due to changes in both the objects and the environment, which remain relatively unexplored. To fill the gap, we propose a time-lapse image dataset inspired by the growth monitoring of strawberries, dubbed “The Growing Strawberries” Dataset (GSD). The data was captured hourly by six cameras, covering a span of 16 months in 2021 and 2022. During this time, it encompassed a total of 24 plants in two separate greenhouses. The changes in appearance, weight, and position during the ripening process, along with variations in the illumination during data collection, distinguish the task from previous MOT research. These practical issues resulted in a drastic performance downgrade in the track identification and association tasks of state-of-the-art MOT algorithms. We believe “The Growing Strawberries” will provide a platform for evaluating such long-term MOT tasks and inspire future research. The dataset is available at <https://doi.org/10.4121/e3b31ece-cc88-4638-be10-8ccdd4c5f2f7.v1>.

### 3.1 Introduction

Multiple Object Tracking (MOT) is an exciting Computer Vision topic with wide applications in autonomous driving [99, 162], traffic monitoring [122, 149], video surveillance [132, 170], etc. While these studies mainly focused on video clips of a few minutes or even shorter [65, 88, 174, 286], consistent tracking over a longer period also has significant implications in real-world contexts. The supervision of cultivation and livestock [84, 91, 223, 285], the progression assessment of lesions and wounds [21, 39, 41, 117, 251], and the microscopic scrutiny of cells [13, 173] serve as intriguing illustrations of this pragmatic scenario. However, there is a lack of research on MOT algorithms applied for long-term purposes, particularly when the intrinsic properties of objects are also simultaneously developing. Furthermore, using a lower capture frequency over extended periods [41, 223, 285] leads to a substantial information loss, thereby heightening the challenges in accurate object tracking. Therefore, there is a pronounced need for a realistic dataset to bridge the gap between current MOT algorithms and their effective application over prolonged periods, so as to facilitate the advancement of effective methods.

The tracking of biological development processes exemplifies a prominent long-term MOT challenge within this particular context [163, 164, 282, 290]. For instance, accurate growth monitoring of fruits and vegetables over time is a key ingredient to successful horticulture. Recent studies have demonstrated that images are feasible non-destructive tools to evaluate the status and quality of fruits [83, 103, 295]. Keeping track of the growth helps in planning harvest schedules, so as to achieve the peak quality and nutritional value of crops. To follow the growth of individual fruits through visual observations, automated image processing is required. We chose strawberries for our research because their 3–7 day life cycle allows for tracking noticeable appearance changes while maintaining a reasonable frame rate. In addition, the natural growth and horticulture activities also introduce object movements along frames.

In this chapter, we introduce the first in-the-wild biological development monitoring dataset, *The Growing Strawberries Dataset (GSD)*. The videos of GSD consist of time-lapse images of strawberry cultivation in six spots at two different greenhouses during the grow-

ing season in 2021 and 2022 respectively. The longitudinal observations of strawberries over their growth are supportive for ripeness assessment, yield prediction, and harvest planning for efficient supply [37, 203]. Unlike the trajectory tracking of common moving objects, *GSD* involves *long-term* tracking of developing objects under a low frame rate, which introduces the two following unique challenges to the MOT task.



Figure 3.1: An example subsequence of image segments from *GSD*, depicting the growth over five days. We can notice dramatic appearance changes and gradual enlargement during the development. In addition, even though the segments are selected to minimize lightness variations, slight differences in segment brightness may still be discernible due to the shifting angles of sunlight.

**Appearance changes** result from the biological growth of strawberries and include changes in color, shape, and size, as depicted in Figure 3.1. These are common properties when a biological object is developing over time, yet limited MOT research has taken these issues into concern. Unlike pedestrians or vehicles that remain visually consistent throughout short videos, strawberries undergo continuous changes in appearance during long-term tracking. Additionally, the visual appearances of strawberries are more similar to each other than those of the traffic participants, which are more colorful and varied. The in-frame discrimination and across-frame association result in challenges for the appearance descriptors, particularly when also confronting dynamic lighting situations and overnight intervals.

**Irregular movements** can be caused by horticulture operations or other human activities. They exhibit occasional co-occurrence with the strawberries’ incremental movements from natural weight increase. For example, the natural increase in fruit weight or deliberate repositioning by horticulturalists can lead to changes in fruit positions. Human intervention can introduce unexpected occurrences like sudden object movements or camera view occlusions. Additionally, harvested fruits may permanently vanish from sight. Since the data is captured hourly, movements could lead to abrupt changes, e.g. position jumps or switches, which make many location changes of *GSD* objects non-linear and irregular. This characteristic from practice calls for research of discontinuous or interrupted videos, which has not been thoroughly investigated, whilst the joint effect with the appearance change still calls for more effective MOT solutions.

The main contributions of our work in this chapter are:

- We established *GSD*, a long-term MOT dataset that used six cameras to track the growth of 12 plants of strawberries in 2021 and 2022 in two different greenhouses.
- We quantitatively compared *GSD* with one popular MOT dataset, *MOT20*, and proposed a unique MOT scenario: the temporal tracking of biologically developing objects in a sparse and long-term data collection.
- We benchmarked the performance of five MOT algorithms to prove the challenges brought by our proposed scenario.



- We visualized the importance of *GSD* from a realistic perspective. In all, our results evidence the limitations of state-of-the-art MOT algorithms for such a long-term MOT task, which highlights the emergence and necessity of proposing *GSD*.

## 3.2 Related Work

In this section, we briefly review popular object-tracking and temporal datasets that promote algorithm development and their limitations on scenarios, in order to highlight the uniqueness and importance of the *GSD*. Secondly, we summarize the concepts of state-of-the-art MOT algorithms and explain our method for evaluating the *GSD*.

3

### 3.2.1 Image datasets for multiple-object tracking

Datasets for MOT predominantly focus on trajectory tracking. Many of the recent tasks of the *MOT Challenge* [179] are motivated by surveillance and autonomous driving. Thus, they mostly focus on the tracking of pedestrians, vehicles, passengers, etc. [79, 88, 174]. For instance, *MOT20* [67] is a widely-used and representative MOT dataset and is extensively utilized by various algorithms as a benchmark to assess their performance. The majority of the sequences are short videos with 10-30 frames per second and lasting for a few minutes [65]. New challenges mainly originate from a higher amount and density of objects in emerging datasets [67, 239, 263]. However, the characteristics of these scenarios are limited. For instance, popular research objects such as pedestrians or vehicles are often characterized by regular or predictable movement patterns. As a result, a greater diversity of datasets is required to facilitate the generalization of MOT in broader domains [65, 286].

The majority of long-term temporal image datasets are used for substantial-scale change detection, e.g. the progress monitoring of construction, deforestation, urbanization, or animal migrations [75, 185, 193, 247, 280]. One of the shared goals is to track the temporal changes of large and (mostly-)static objects or of a comprehensive overview of objects. Therefore, the main concern in these studies is the pattern differences across images. On the other hand, these datasets have limited potential to motivate the development of MOT algorithms due to the restricted spatial movements of objects.

### 3.2.2 Image datasets for plant science

Image datasets are vital for plant science. Sequential images are a practical data type to accomplish non-destructive tests and continuous growth monitoring. The majority of plant science research involving non-destructive testing of images is carried out within controlled and calibrated laboratory settings [165, 196, 295]. However, for fruits that do not ripen after harvest, it becomes impractical to rely on lab data for recording status updates during their growth. Existing in-field datasets primarily focus on one-shot fruit detection and lack information on the ripening progress due to limited object appearance changes over a short period [86, 140, 199, 292]. Moreover, hyper-spectrum images (HSI) play a valuable role in plant studies by developing numerical indicators and training machine-learning models [86, 114, 295]. Yet, integrating these images into agriculture practices is resource-intensive, given the already costly nature of HSI data collection. Therefore, we advocate for a more practical solution: an integrated temporal dataset merging images in the visual and near-infrared spectrum. The scarcity of non-visual images further emphasizes the need for such a comprehensive dataset.

### 3.2.3 Algorithms for multiple-object tracking

Online MOT algorithms aim to perform real-time tracking of multiple objects in video sequences by continuously updating object identities and associations. Tracking-by-detection is the most widely-used strategy in achieving online MOT [12, 74, 262]. The strategy enhances the algorithms' adaptability and robustness, enabling them to easily accommodate and perform well in diverse scenarios. In addition, it has less reliance on high FPS of data collection than strategies building end-to-end detector-trackers such as [25], which exhibits a higher potential for successful adaptation and utilization in long-term MOT problems. Offline MOT solvers are also powerful tools as they utilize batches of frames [35, 213, 273]. Since the computation effort grows tremendously on larger datasets<sup>1</sup>, it is out of the scope of the context of our dataset. Thus, online MOT algorithms are more applicable in GSD.

Algorithms following the tracking-by-detection strategy consist of two stages: (i) applying object detection models and (ii) associating bbox across frames. Research towards better (near-)real-time performance mainly focuses on enforcing the associating algorithm or a better interconnection between the two stages [262]. Generally, the association step concerns two criteria [239]: **(i) The trajectory and motion of objects.** Many MOT algorithms are developed based on the Simple Online and Real-time Tracking (SORT) algorithm, in which a Kalman filter framework is applied to analyze the velocity vectors [27, 45, 297]. The utilization of inertia measurement is a widely recognized approach for expeditiously handling the MOT task. Nevertheless, researchers argue that trajectories of spatially close objects are difficult to be distinguished [273]. **(ii) The appearance of objects.** Deep learning techniques are usually applied to encode the appearance information of targets [52, 262, 273, 276]. Field-specific object properties are often integrated to enhance association performance [41, 213]. Particularly, when the frames are discontinuous or when the objects are occluded, appearance features are crucial in re-identifying and associating the tracklets to achieve consistent global tracking [229, 298, 301–303]. Nevertheless, the sparsity of the image collection for GSD indicates a longer interval between frames, which exacerbates the existing complexity of the task.

## 3.3 The Growing Strawberries Dataset (GSD)

We aimed to establish a dataset about prolonged object tracking in a real-world setting for the purpose of long-term MOT. The growth of strawberries is a good example of a natural biological development process. Appearance changes and irregular movements happen during this ripening process. Such dynamics reveal special characteristics that are also shared among all kinds of agricultural crops.

To this end, we used six cameras (three RGB + three OCN<sup>2</sup>) to track the growth of 12 *Favori* plants over 30 weeks in 2021 and 32 weeks in 2022, in two greenhouses with different cultivation setups in The Netherlands. The cameras were paired in three sets, denoted as RGB/OCN-1/2/3. They captured time-lapse images in the greenhouse, such that videos of the entire ripening process were archived. We provide human-annotated bboxes for every strawberry, at all growth stages, and identities for corresponding trajectories.

<sup>1</sup>An example on GSD is demonstrated in the supplementary materials.

<sup>2</sup>The channels are: Orange/615nm, Cyan/490nm, Near-Infrared/808nm.

### 3.3.1 Data collection setup

Since the ripening lasts around 7-14 days, we used hourly images for growth monitoring, such that a complete track of the plant is ensured with circa 100 observations. The strawberries were cultivated in planting baskets that hung from the ceiling. A heating pipe was hung beneath each planting basket. The cameras were attached to the heating pipe on the opposite side of the strawberry plant. Figure 3.2 illustrates the detailed setup of the cameras in the greenhouse.

Both cameras faced the plants from parallel perspectives, where the OCN images were taken with a large view overlap with the RGB ones to provide hyper-spectral information. On average, 28 strawberry fruits from 4 plants were monitored by one RGB camera. We index all three RGB cameras as *RGB-1,2,3*. Figure 3.3 shows the annotations of an example image sequence taken by camera *RGB-3*.

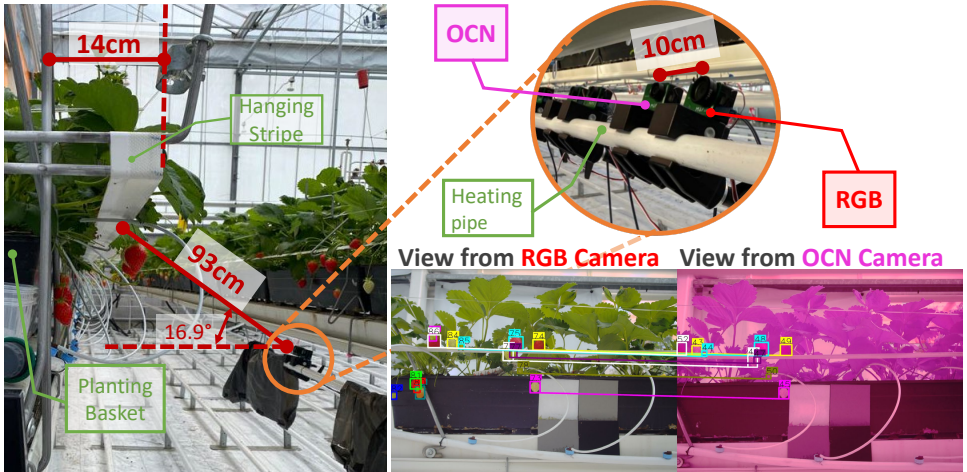


Figure 3.2: Detailed setup in the greenhouse in 2021. The left photo shows the positions of the white stripes, the planting baskets, and the heating pipe, which were all placed in parallel. The distance from the edge of the white stripes to the camera lens was 93 cm. The RGB camera was placed 10 cm to the left of the OCN camera of each pair, as shown on the top right. The elevation angle of both cameras is  $16.9^\circ$ . Sample images from *RGB-1* and *OCN-1* are shown in the bottom right. Identical strawberries are color-coded. The setup is similar in 2022 with slightly varied dimensions.

### 3.3.2 Ground-truth annotation

The trajectory annotations of the strawberries consist of bboxes with track identifiers (track IDs). To remain consistent in labels, the first round of annotation was performed by a single person. Subsequently, two separate reviewers performed a manual check on the annotations to ensure accuracy and to mitigate potential labeling errors or personal biases. In this way, we guarantee accurate annotations. Examples of the MOT annotations are depicted in Figure 3.3 and Figure 2.2.

All the images are  $4000 \times 3000$  pixels. Due to the continuous data collection spanning the entire day, the illumination conditions exhibited significant and periodic variations.

We therefore set up a brightness threshold and defined a subset specifically for the following benchmarking experiments.

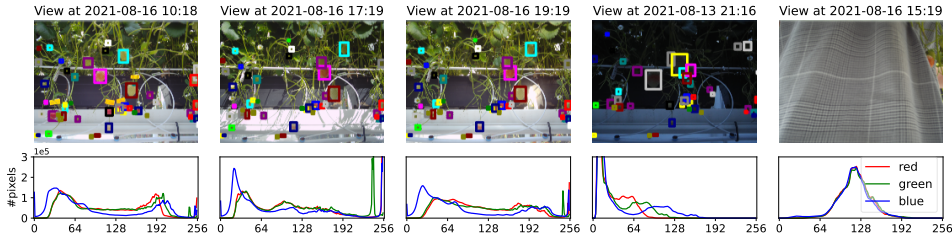


Figure 3.3: The upper row lists five image samples taken by *RGB-3*. The capture time is indicated in the title. The 1<sup>st</sup> - 4<sup>th</sup> images depict the normal changing pattern of sunlight during the day. We use the 4<sup>th</sup> image from a different date because the dawn and dusk were not captured every day. The 5<sup>th</sup> image shows how the view might be blocked due to human activities. The plot beneath each photo is the corresponding color spectrum from R/G/B channels respectively. The x-axes indicate the color value (encoded as 0-255). The y-axes are the power of the color spectrum with a shared amplitude. The color-coded rectangles illustrate the ground-truth (GT) bbox and trajectory annotations.

**Day images.** The RGB images that were taken under normal lighting conditions are the majority share of the growth tracking task. Examples are as depicted in the left three photos in Figure 3.3. We call this subset the “day images”. Quantitatively, they were defined as the images with luminance<sup>3</sup> higher than 50. As is illustrated by the first three columns in Figure 3.3, when the zenith angle of the sun changes during the day, the color spectrum of the photo shifts. This is a practical challenge brought by the in-the-wild data collection, which also aggravates the variation of object appearances.

**Remainder images.** The annotations are available for all frames until most strawberries became invisible when the view became very dark or when the camera was occluded by human activities (e.g. the 5<sup>th</sup> photo Figure 3.3). We defined the subset “darker images” as the photos that were taken under insufficient daylight (i.e. image luminance  $\leq 50$ ) but the strawberries were still visible to be annotated, for example, the 4<sup>th</sup> photo in Figure 3.3. Nevertheless, without additional brightness normalization, darker images degraded the performance of the detection models. Considering that the number of darker images was limited (at most once during dawn and/or dusk), we excluded them in the benchmarking experiments to keep a fair performance comparison.

**Trajectory annotations.** Overall, the trajectories of strawberries have an average length of 152 bboxes, yet it ranges from 2 to 600+ bboxes. The extra-long tracks resulted from slower growths under cool temperatures. In fact, there is still a notable proportion of tracks that last less than 20 segments, which are mostly incompatible with the natural growth cycle of strawberries. Two major reasons for these short tracks are: (i) the back-layer ones only started to be visible after re-position practices from humans because the strawberries grew in dense clusters; (ii) the growths were only partially monitored because the size increases of strawberries might squeeze the others out or into the frames.

<sup>3</sup>Luma, calculated according to ITU-R BT.601 standard [29].

Table 3.1: Statistical overview of the RGB images of *GSD*. The 2<sup>nd</sup> column lists the duration of data collection. The 3<sup>rd</sup> and 4<sup>th</sup> columns note the amounts of all images and the images used in the benchmarking studies respectively. The last two columns present the total number of bboxes and trajectories. An overview of the OCN images is presented in the supplementary materials.

Camera	Period	Total img	Anno. img	Total bbox	Total tracks
RGB-1	Apr 23 - Nov 9, 2021	4786	2823	67957	492
RGB-2	Apr 23 - Nov 9, 2021	4785	2638	64434	392
RGB-3	Jun 29 - Nov 9, 2021	3181	1761	70641	431
RGB-1	Feb 22 - Oct 3, 2022	5128	3369	93439	540
RGB-2	Feb 22 - Oct 3, 2022	4699	3062	117291	872
RGB-3	Feb 22 - Oct 3, 2022	5156	3330	109946	754

### 3.3.3 Data characterization

Compared to pedestrian-focused datasets such as the *MOT20*, *GSD* objects usually are more similar looking to one another, whilst they have more evident appearance changes over frames. In addition, larger and more irregular movements are observed in *GSD* trajectories.

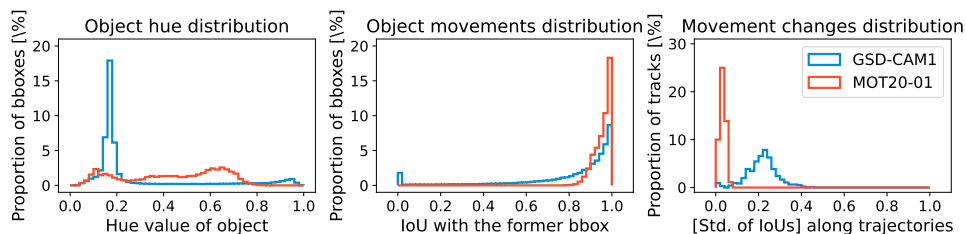


Figure 3.4: Quantitative comparisons of *GSD-2021-RGB-1* and *MOT20-01*, using the GT annotations. The 1<sup>st</sup> spectrum shows the distribution of object colors, posed by the average hue from the center 50% area of the bboxes. The 2<sup>nd</sup> plot illustrates the distribution of overall object movements, using IoU as the metric. The 3<sup>rd</sup> plot presents the standard deviations of the bbox IoU of each trajectory.

Figure 3.4 presents comparisons of the color and movement distribution of the sequence *GSD-2021-RGB-1* (shortened as *RGB-1* in the following text) and *MOT20-01*. The 1<sup>st</sup> subplot shows the hue value, calculated from the HSV color space [159], of all bboxes. Here, *RGB-1* shows a higher degree of monotonicity among the observations compared to *MOT20-01*, which also indicates larger challenges to the feature extractors. Nevertheless, for the same *GSD* object, the color keeps changing due to its biological development over the time span, together with the illumination condition. An example is shown in Figure 3.5.

We measured the object movements by the Intersection of Union (IoU) of observations in adjacent frames because a large proportion of MOT algorithms consider a sequential matching of objects by including more and more frames in analysis. Followingly, larger movements are indicated from the left of the x-axis in the 2<sup>nd</sup> subplot. As Figure 3.4 shows, the movements of *GSD* objects are more spread out, whilst the *MOT20* objects exhibit

slower movements, holding a minimum IoU of 0.8. Moreover, there are a few bboxes that have minimal intersections with its previous observation, which introduces extraordinary challenges to the inertia measurement and the association algorithms. We also calculated the standard deviation (std.) of the IoU of each trajectory. The value indicates the irregularity of how each object moves. As the 3<sup>rd</sup> subplot shows, such irregularity in *RGB-1* is higher in magnitude.

## 3.4 Benchmark Studies

Since *GSD* has a large number of high-resolution images, we primarily restricted our attention to lighter, online solvers. In addition, we applied *GMTracker* [106] on a small subset<sup>4</sup> to exemplify the performance with an offline solver.

We assessed the performance of the four MOT algorithms to demonstrate the challenges presented by *GSD*: (i) *ByteTrack* [297] that performs an Intersection of Union (IoU) analysis after applying the Kalman filter as SORT does; (ii) *Observation-Centric SORT* (OC-SORT) [45] that is enhanced against noised and non-linear movements; (iii) *DeepSORT* [276] that introduces appearance descriptions to identify objects before applying the matching by movements; (iv) *StrongSORT* [74] that improves the movement measurement and its balance with the appearance features. On top of the original settings, we altered the appearance-cost parameter ( $\lambda$ ) of StrongSORT to introduce different emphases for appearance and motion information in the association stage.

Since all the algorithms share the tracking-by-detection strategy, we present our experiments from three aspects: the overall MOT performance of the algorithms (and variations), detection-stage impact, and tracking-stage influence. Drawing upon the results, we explore the potential implications stemming from the distinctive characteristics of the *GSD*, which we contend represent challenges within biological development tracking applications.

### 3.4.1 Application of MOT algorithms on GSD

By dividing the subsets by cameras, we first trained three YOLOX-x models with a “leave-one-camera-out” cross-validation strategy. We employed the detections on the test set for the MOT performance evaluation. We conducted all experiments using the daytime subset of *GSD-2021* to ensure that darkness-related distractions were avoided, thus enabling a more equitable comparison. We reduced the IoU threshold to 0.1 in the association stages, due to the different object movement patterns as indicated in Figure 3.4. We independently developed autoencoders to serve as the appearance descriptors for DeepSORT and StrongSORT. Detailed parameter settings and searching are noted in the supplementary.

We evaluated the overall performance by the widely-known MOT criteria: the Higher Order Tracking Accuracy (HOTA) [161] and the Multi-Object Tracking Accuracy (MOTA) [26]. The performance of track identification is described by accuracy (AssA), recall (AssRe), precision (AssPr), and the balanced criterion IDF1 [211]. We counted the number of identity switches (IDS) and the interruptions of trajectories (Fragmentation/FM) and divided the values by the amount of ground-truth (GT) tracks to compare with other datasets, e.g. *MOT20* or *MOT17*. They are noted as “IDS/Tr” and “FM/Tr” respectively.

<sup>4</sup>Our justification for using the subset is provided in the appendix.



### 3.4.2 Assessing comprehensive MOT performance

The performance metrics are summarized in Table 3.2. In general, the algorithms exhibited inferior performances on *GSD-2021* compared to their achievements on *MOT20*. Notably, compared to more comprehensive metrics such as HOTA and MOTA, all the criteria related to the evaluation of bbox association and trajectory identifications, e.g. IDF1 and AssA, indicate intense performance drops from their original benchmarks. The performance downgrade came with exaggerated frequencies of ID switches and trajectory interruptions. The numbers suggest that the *GSD* tracks have a relatively higher discontinuity as per the MOT algorithms, which could be caused by the increasing changes during the prolonged data collection. The results further evidence that *GSD* introduces a more challenging task than *MOT20* for the state-of-the-art MOT methods.

As shown in Table 3.2, ByteTrack performed the best in terms of HOTA, and OC-SORT was better in limiting the switching of track IDs. When adjusting the parameter  $\lambda$  in StrongSORT to increase the emphasis on motion over appearance matching, notable improvements in overall performance were observed. Hence, associating bounding boxes based on inertia measurements is proved to be relatively more applicable in this case. Nevertheless, we also notice that, whilst shifting the focus to object movements lessened the IDS/Tr, it also led to higher FM/Tr. It indicates that the current appearance-based methods need to be improved to handle data collected at such a sparse frequency.

Upon a dedicated processing time of 112 hours, GMTracker associated the first 750 frames of *RGB-1*. Notably, apart from the training process that already required substantial time and computational memory resources, it devoted over 2 hours to processing some of the frames, with a maximum time of 7498 seconds for a single frame. As evident in Table 3.2, this end-to-end network's performance matched the other benchmarks, yet was achieved by significantly more intensive use of resources. Hence, we remain our focus on the lighter solvers in subsequent discussions.

### 3.4.3 Detection performance and impact

To verify the attainable optimal solution of the object-detection stage, we evaluated two state-of-the-art object detection methods on *GSD*, the anchor-based detector Faster R-CNN and the anchor-free detector YOLOX-x, following the “leave-one-camera-out” strategy. The Average Precision (AP) obtained by both models is noted in Table 3.3.

Due to limitations from the volume and properties of the training data, the detection performances were not so competitive as the private models that were specifically trained for the pedestrian-tracking challenges [74]. However, under a single-category setting, both detectors' performances aligned with the published detections of the *MOT20* testing set [67] and their respective model developers' benchmarks [87, 278]. Although these performances are not directly comparable due to the differences in the validating datasets, we argue that the difficulty level of the object detection task on *GSD* is not significantly higher than other datasets. Therefore, the main challenge brought by *GSD* lies in the association stage, which is also the main task of MOT.

Moreover, for a fair comparison of algorithm performances on *GSD*, we also present the metrics obtained from the public *MOT20* detection sets, as provided on the MOTChallenge website [128, 191]. As shown in Table 3.2, the MOTA scores achieved using the public *MOT20* detections are even lower than the results obtained on *GSD*. This divergence can

Table 3.2: Performance metrics of the four original and two tailored MOT algorithms on the daytime subset of *GSD-2021* (\*the GMTracker was only applied on the first 750 frames of *GSD-2021-RGB-I*). The results are compared with the performance metrics of the same algorithms implemented on the MOT20 test set, using the results with private detections in [45] and [74] and the results with public detections on the MOT20 challenge website. (\*\*The performance of GMTracker was compared with its results on the *MOT17* test set, using the metrics claimed by [106]). The differences are indicated by red and teal texts that are noted at the top right of each value, representing performance degrades and improves, respectively. 'Pvt' and 'Pub' indicate whether the gap is with benchmarks using the private or public detections (and encoders if applicable). If one value is shown, it is compared with only the metrics claimed in the paper, obtained from private detections (and encoders). In terms of StrongSORT,  $\lambda$  is the default weight on the appearance cost, and  $\lambda'$  indicates an altered value.

MOT Algorithm	HOTA	MOTA	IDF1	AssA	AssRe	AssPr	IDS/Tr	FM/Tr
ByteTrack	39.8 <sup>pvt:-21.5</sup> pub:-16.6	70.7 <sup>pvt:-7.1</sup> pub:+3.7	39.4 <sup>pvt:-35.8</sup> pub:-30.8	25.6	29.3	70.0	5.2 <sup>pvt:+4.2</sup> pub:+4.7	5.4 <sup>pvt:+4.2</sup> pub:+4.0
OC-SORT	39.7 <sup>pvt:-22.4</sup> pub:-14.6	68.5 <sup>pvt:-7.0</sup> pub:+8.6	39.5 <sup>pvt:-36.4</sup> pub:-27.5	25.9	29.4	72.5	4.5 <sup>pvt:+3.8</sup> pub:+4.1	5.3 <sup>pvt:+4.4</sup> pub:+3.4
DeepSORT	34.5 <sup>-22.6</sup>	49.3 <sup>-22.5</sup>	33.0 <sup>-36.6</sup>	22.3	26.4	62.3	8.4 <sup>+7.3</sup>	5.4
StrongSORT( $\lambda=0.98$ )	36.1 <sup>-25.4</sup>	49.3 <sup>-22.9</sup>	34.0 <sup>-41.9</sup>	23.9	27.7	64.7	8.8 <sup>+7.9</sup>	5.1
StrongSORT( $\lambda'=0.5$ )	38.5	59.9	35.8	25.4	27.9	76.8	6.2	5.8
StrongSORT( $\lambda'=0.02$ )	38.6	60.4	35.8	25.5	27.9	77.7	6.0	5.9
GMTracker*	37.7	60.2 <sup>+4.0**</sup>	31.7 <sup>-32.1**</sup>	22.2	23.2	85.0	20.3 <sup>+19.6**</sup>	3.8



be attributed to the limited accuracy of the public detection set. Nevertheless, even when emphasizing track identification metrics like HOTA and IDF1, substantial differences persist. Additionally, the algorithms' IDS/Tr and FM/Tr on *GSD* are still significantly higher compared to those on *MOT20*.

Table 3.3: The first three rows show the AP of the detections of *GSD* and the public *MOT20* detections. All values are averaged over the three test sets split by the “leave-one-camera-out” strategy. The latter two rows present the original mAP benchmark for comparison.

Model-Dataset Configuration	AP
YOLOX-x on <i>GSD</i>	55.7
Faster R-CNN on <i>GSD</i>	55.8
Faster R-CNN on <i>MOT20</i> [67]	57.6
Faster R-CNN on COCO [278]	40.2 (mAP)
YOLOX-x on COCO [87]	59.2 (mAP)

### 3.4.4 Association performance and impact

To compare the specific accuracy of track association, we decouple the detection performance by benchmarking StrongSORT on the ground-truth detection annotations from *GSD-2021* and *MOT20*. For validation, we used *RGB-1* and *MOT20-01* as examples. As shown in Table 3.4, both MOTA were boosted due to the perfect-detection assumption. However, the improvements in HOTA and IDF1 on *RGB-1* experiment were not so significant as those in the *MOT20-01* experiment. Furthermore, noticeable gaps in performance are observed in IDS/Tr and FM/Tr.

The influence of the parameter  $\lambda$  follows a similar pattern as previously described – the emphasis on motion or appearance results in a trade-off between IDS/Tr and FM/Tr. Referring to the data characterization, the higher similarity in appearances among the *GSD* objects and the dynamic variation of them may contribute to the downgraded IDS/Tr performance. Considering that the data was collected over prolonged periods, the incorporation of appearance features is expected to assist in consolidating the fragmented tracklets, e.g. after human activity or overnight. Hence, it is advisable to tailor the utilization of appearance matching in MOT algorithms for scenarios involving sparse frame rates.

### 3.4.5 Evaluating results from a downstream application

One contribution of *GSD* is its provision of valuable information for agriculture practices, enabling precise anticipation of crop growth. Since the natural ripening pattern of strawberries is growing from green to red, we utilized the  $A^*$  channel from the CIELAB color space [221], which essentially represents the levels of green or magenta. In Figure 3.5, the blue curve demonstrates a sample  $A^*$  variation of the object across frames. Marking associated observations with colored dots and un-associated ones with empty dots, the depicted process is fragmented into five segments by four tracklets suggested by ByteTrack (for obtaining the highest HOTA in Table 3.2), involving two IDS in tracklet #21 and #40.

Table 3.4: Association-stage performance comparison of StrongSORT, with variations on the appearance-cost parameter  $\lambda$ , applied on the ground-truth detections. We use  $G$  and  $M$  to represent  $GSD-RGB-I$  and  $MOT20-01$  respectively in this table. In all experiments, the ground-truth locations of the bboxes were used, such that the algorithm performance was not influenced by the detection accuracy.

Algorithm	HOTA		MOTA		IDF1		IDS/Tr		FM/Tr	
	GSD	MOT20	GSD	MOT20	GSD	MOT20	GSD	MOT20	GSD	MOT20
StrongSORT( $\lambda=0.98$ )	51.5	98.6	83.5	99.5	43.5	98.6	6.3	0.0	3.7	0.0
StrongSORT( $\lambda'=0.5$ )	51.4	99.3	83.6	99.5	42.5	99.4	5.9	0.0	4.6	0.0
StrongSORT( $\lambda'=0.02$ )	52.2	99.2	85.1	99.5	42.7	99.4	5.2	0.0	4.2	0.0

Notably, during the crucial period when the strawberry underwent the transition from green to red, which is a crucial factor in determining the timing of harvest, ByteTrack was unable to provide a thorough description of this transformation.

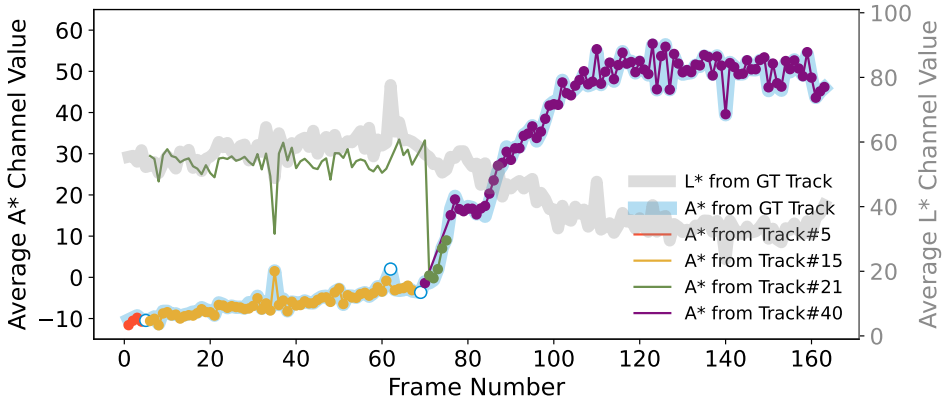


Figure 3.5: The color change of an example strawberry under the GT trajectory and the ByteTrack results. The x-axis indicates the sequence of frames. The y-axes are for the average  $A^*$  values (scales on the left) and  $L^*$  values (scales on the right) of the observations. The blue and gray translucent strokes illustrate the value of the GT annotations. The lines with filled dots are identified observations by ByteTrack, which are color-coded to indicate each track ID. If the object in one frame is not associated with any of the tracks, we put an empty dot on the  $A^*$  curve from GT.

To evaluate the significance of performance deficiency from the perspective of realistic, downstream applications, i.e. tracking the biological development of objects, we set up thresholds to define the “cherry-picked tracks” that record relatively comprehensive monitoring of growth patterns. We chose tracks based on more significant variations of the object’s transition from green to red, determined by the changes in the  $A^*$  channel values in the CIELAB color space, or simply select the tracks with longer lengths. These tracks were considered “more important” ones as they provide more complete information about the growing progress of the crop. We implemented incremental thresholds to perform stricter filtering of their importance.

Figure 3.6 discusses the specific performance of ByteTrack, the relatively more capable solution for *GSD*, on the different filtered subsets of *RGB-1*. As is depicted, the recall of track association declined as the track became more comprehensive about the biological development cycle. Simultaneously, there were increases in *IDS/Tr* and *FM/Tr*. The track length played a more significant role in the deterioration of performance under this particular scenario.

Viewing from an application-oriented standpoint, the growth-tracking task also targets monitoring pivotal stages when fruits are ripening swiftly. Therefore, it is argued that there is potential for advancing state-of-the-art MOT algorithms, particularly in accurately identifying and associating objects within similar biological development processes.

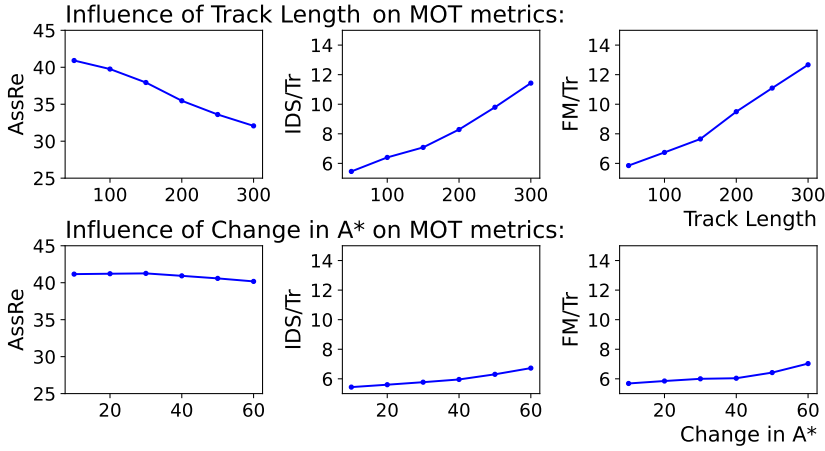


Figure 3.6: MOT performance change by selection criteria of trajectory subsets, demonstrated by recall (1<sup>st</sup> column), ID switching (2<sup>nd</sup> col.), and times of fragments (3<sup>rd</sup> col.) of tracklets. The first row illustrates the impact on the performance metrics when different minimum lengths filtered the tracks. Experiments for the second row selected the tracks based on the differences in the average  $A^*$  value of the last three and the first three bboxes.

### 3.5 Conclusion

With this chapter, we propose a fully-annotated dataset that tracks the growth of in total of 3481 strawberries over 30 weeks in 2021 and 32 weeks in 2022 in two different greenhouses: *The Growing Strawberries Dataset (GSD)*. It reveals a unique Multiple-Object-Tracking (MOT) challenge – following biologically developing instances over a prolonged period. In *GSD*, progressive appearance change and irregular movements are captured from the longitudinal observations of cultivation practices. For example, human interference with the sparse frame rate introduced drastically non-linear movement, which is challenging for many algorithms.

We benchmarked the performance of four online MOT algorithms on *GSD*. The obtained result metrics highlight the need for advancing MOT methods, particularly in associating the bounding-box association for long-term MOT tasks. The tracking continuity was affected by both appearance changes and diverse object motions, which also presented a trade-off when fine-tuning StrongSORT. Furthermore, an offline algorithm demonstrated the computational effort required to handle a large dataset such as *GSD*, yet achieving similar metrics. In summary, the results call for algorithms that could improve track associations while utilizing the features properly and efficiently.

Essentially, biological development is the principal property that makes the *GSD* challenge unique, but it can also provide insights for other long-term MOT tasks. For instance, monitoring other processes with incremental changes, such as cellular growth and corrosion expansion, etc. The information provided by more than the visual spectrum is also supportive of plant science [37, 203]. The *GSD* challenge highlights the need for reliable methods to handle in-the-wild data imperfections. The inevitable real-world challenges point out potential future research for robust data utilization.



## 4

# Predicting Sugariness Using Non-Destructive and Affordable Hardware

## 4

This chapter explores the results of a conceptual methodology for training machine learning models on various data types. We leverage in-field data to make accurate predictions about a strawberry's sugariness, a crucial characteristic of interest for supply chains that have traditionally been assessed through destructive testing or laboratory analysis after harvest. While there exist some overlaps with Section 2.2, Section 4.2 focuses on the specific input data and labels used in this research.

This chapter was previously published as: Junhan Wen, T. Abeel, M. de Weerd. 2023. " "How sweet are your strawberries?" : Predicting sugariness using non-destructive and affordable hardware." *Frontiers in Plant Science* 14 (2023): 1160645. [[264](#)]

*Global soft fruit supply chains rely on trustworthy descriptions of product quality. However, crucial criteria such as sweetness and firmness cannot be accurately established without destroying the fruit. Since traditional alternatives are subjective assessments by human experts, it is desirable to obtain quality estimations in a consistent and non-destructive manner. The majority of research on fruit quality measurements analyzed fruits in the lab with uniform data collection. However, it is laborious and expensive to scale up to the level of the whole yield, and such a “harvest-first, analysis-second” method also comes too late to decide to adjust harvesting schedules.*

*In this research, we validated our hypothesis of using in-field data acquired via commodity hardware to obtain acceptable accuracies. The primary instance that the research concerns is the sugariness of strawberries, described by the juice’s total soluble solid (TSS) content (measured in “Brix or shortened as Brix). We benchmarked the accuracy of strawberry Brix prediction using convolutional neural networks (CNN), variational autoencoders (VAE), principal component analysis (PCA), kernelized ridge regression (KRR), and support vector regression (SVR), based on fusions of image data, environmental records, and plant load information, etc. Our results suggest that: (i) models trained by environment and plant load data can perform reliable prediction of aggregated Brix values, with the lowest RMSE at 0.59; (ii) using image data can further supplement the Brix predictions of individual fruits from (i), from 1.27 to 1.10, but they by themselves are not sufficiently reliable.*

## 4

## 4.1 Introduction

Soft fruits such as strawberries, raspberries, blueberries, etc. are popular and profitable fruit varieties. The annual consumption of strawberries in Europe is estimated to be more than 1.2 million tonnes, which leads the market share of horticultural crops [44, 47, 48]. Worldwide production of strawberries is stable with increasing demands and prices and is continuously growing even through the COVID-19 pandemic [34, 48, 49, 233]. However, without the protection of hard skins, soft fruits are vulnerable during production and post-harvest activities. This results in significant food waste and economic loss [100, 208, 237]. The food loss and waste comprise up to 50% loss along the supply chain in some countries [130, 205], among which the production loss is the majority, which consists of up to 20% [201, 244]. It has been estimated that for every ton of food waste, €1,900 of production and processing costs are lost. Moreover, it is argued that 50% of the waste could be edible [237].

The nutritional and economic value of crops is influenced by the harvesting strategy. However, subjective assessments and inappropriate maintenance of fruit quality could bring conflicts in logistics planning between suppliers and distributors, which results in even further post-harvest loss [77, 204]. Therefore, early decision-making supports both ecological and economic interests. To make logistic and harvesting decisions as early as possible, it is highly desirable to predict the quality of ready-to-harvest strawberries in the field [1, 60, 147, 235].

Multiple variables determine the quality of a strawberry, including maturity, shape, sweetness, and firmness [153, 178, 279]. As the majority of strawberry products are consumed fresh, the taste is the highest priority for most European consumers of strawberries [48, 49]. Therefore, we narrow our research scope of this chapter to concern the

interior quality of the fruit, which is not directly told by their appearances: this study explores the assessment of the level of sweetness of strawberries, which is quantitatively described by total soluble solid (TSS) content in the juice of freshly harvested fruits, using informatics and machine learning (ML) approaches.

Traditionally, the TSS content is measured by a refractometer and expressed in degrees Brix (°Brix), often shortened to simply “Brix” [18]. The measurement is expensive in both labor cost and capital because the samples that are sent to destructive measurements can no longer be sold [6, 96]. To reduce errors and optimize the supply chain, there is a desire for more accurate, quantitative, and non-destructive tools to assess the quality of each fruit [166, 253]. Therefore, we explore the feasibility of Brix prediction with easily-acquirable data, such that the prediction can be carried out on-site without specific fruit preparation.

Related research has demonstrated the applicability of computer vision (CV) in grading the quality of fruits [135, 154, 183, 294] and in assessing specific quality attributes [2, 18, 178, 250]. CV and spectral analysis with hyperspectral imaging (HSI) are popular techniques that are often used to investigate intrinsic properties [6, 9, 85, 158]. High prediction accuracy was achieved when fruit photos were acquired under a (mostly-)uniform experiment setup [166, 186, 227, 274, 279]. Such a setup requires delicate devices that, on the other hand, limit further development for real-world applications and large-scale sampling. Moreover, the “harvest first, analysis second” methodology limits the possibility of adjusting the harvest strategy for supply chain optimizations because strawberries stop growing after being harvested. Hence, our study concerns the implication of the fruit’s intrinsic characteristics by its appearance under natural light, when the fruit is still on the plant.

Meanwhile, the micro-climate in the greenhouse and the horticultural treatments strongly influence the harvest quality and pace of growing [56, 70, 232]. The temperature, humidity, CO<sub>2</sub> level, lighting conditions, and irrigation are proven to be crucial factors [15, 60, 110, 181, 232]. The crop load is also argued to influence the quality of fruits [23, 61, 256]. In modern horticulture, environmental data is readily collected by field sensors or climate computers in most greenhouses [105, 181, 219, 232]. Nevertheless, these point measurements cannot provide distinctive information to specify the quality of individual fruits. Thus, our research introduces approaches to integrate in-the-wild fruit images with environmental and plant-load data in predicting the Brix values of individual fruits.

By investigating the performances of Brix prediction models, we aim at providing insights into two main questions: (1) how accurately can the models estimate the Brix values by different sets of inputs? and (2) which data are valuable for training the Brix prediction models? The research addresses these questions and contributes from four perspectives: (i) we collected and labeled a dataset of strawberry images and quality measurements, using commodity hardware; (ii) we designed a conceptual methodology of non-destructive quality estimation; (iii) we shaped and implemented our methodology to predict the strawberry sugariness; (iv) by comparing the model performances, we suggest how to develop reliable prediction models by CV and ML methods.



## 4.2 Materials and Methods

### 4.2.1 Dataset

Data were collected from May 2021 to November 2021. This was carried out on overwintered trays of *Favori* strawberry plants in a greenhouse at the Delphy Improvement Centre B.V. (Delphy) in Bleiswijk, the Netherlands. Strawberries were cultivated in baskets that were hung from the ceiling in the greenhouse. For the plants monitored by the cameras, the harvesting frequency is mostly once per week, or twice per week when the strawberries grow faster in warmer periods. There is exactly one harvest round per day, so we use “from a harvest” to describe the data collected from the same date.

The data collection setup consisted of the following parts: (1) static cameras facing the planting baskets to take periodic photos; (2) Brix measurements of the strawberries by the horticulturalists from Delphy; (4) physical labels on the branches to identify the measurement results of a strawberry with its appearance in images; (4) climate sensors to record the environment in the greenhouse and the outside weather; (5) plant loads, represented by the average number of *Favori* fruits and/or flowers per unit area; (6) other logs about the plant cultivation.

Representations of individual strawberries were the major inputs to train the Brix prediction models. We considered image data because they are objective and distinct. The images were collected hourly with a time-lapse setting. The same sections of six example images are shown in Figure 4.1. As is shown in the figures, we stuck a yellow label to indicate the ID of a strawberry a few hours before the harvest (namely the “ID label”), such that the strawberry’s appearance in the images can be connected to the measurement results. The measurement data that are assigned to identified strawberries are called the “connected measurements” in the following text.

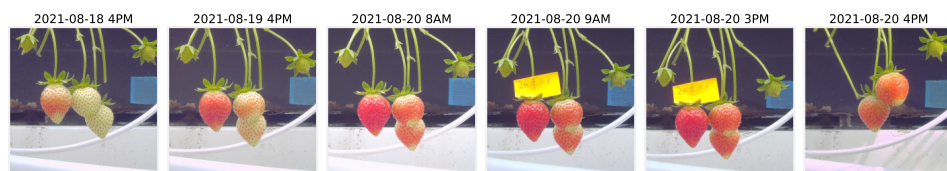


Figure 4.1: Illustration of the time-lapse images, with the same sections selected across six frames. The timestamps of data collection are indicated above the images. As can be seen, by 9 am on 2021-08-20, the yellow physical label was attached to the branch, and strawberry 20.8.1.1 was then harvested between 3 to 4 pm. Thus, the last time when it was observable on images was 3 pm, 2021-08-20.

Based on previous research on influencing factors of strawberry qualities [15, 51, 61] and the expertise of our collaborating horticulturalists, temperature, humidity, radiation level, CO<sub>2</sub> density, and relevant plant treatment records (additional lighting, watering were all considered as the environment data. The number of fruits and/or flowers per unit area was counted weekly and noted as the “plant load”. Both the environment and plant load data were collected by Delphy.

The strawberries with the ID labels were stored separately. On the same day of the harvest, researchers from Delphy measured the Brix value and the firmness of those strawberries, with a refractometer and a penetrometer respectively. The size category is defined

by a ring test, and the ripeness level is evaluated according to the experience of the greenhouse researchers.

## 4.2.2 Methodology

We segmented the strawberries from the in-field images, such that only the pixels that describe the sample strawberry were analyzed. We trained a *Mask R-CNN* model [107] with a *ResNet101* backbone for semantic segmentation. We used the *Detecron2* platform [278] to build the model. The *ResNet101* backbone was pre-trained on the *ImageNet* dataset. We resized the image segments to  $200 \times 200 \times 3$  pixels. They were the raw inputs for Brix prediction and feature extraction in the *image-only experiment*, the *image-with-env experiment*, and the *image-with-Brix experiment*. We considered only the last available observations, e.g. the strawberry segment from the 5<sup>th</sup> image in Figure 4.1. In this way, we limited the quality changes between when it was in the image and when it was measured. We also normalized the colors of the images to reduce the distraction from the changing lighting conditions during the day by applying elastic-net regressions at the red, green, and blue channels respectively.

To analyze the images in the *image-only experiment*, we built convolutional neural networks (CNNs) and variational auto-encoders (VAEs) to analyze and encode the image segments of individual strawberries with multi-layer perceptrons (MLPs). The models were either trained from scratch or with weights pre-trained by other popular datasets such as the *ImageNet* [68]. Details of model architectures can be found in the supplementary materials. We also introduced principal component analysis (PCA) in the experiments for feature dimensionality reduction and model regularization [89, 225]. By taking the largest differences among the pixel data, PCA helps to exclude disturbance from the shared information of strawberry images to some extent. Hereafter, we use the word “encode” to represent the process of dimensionality reduction by the encoder parts of the VAEs and/or PCA. We use “attribute” to describe the content of information that our model concerns. “Feature” or “input” represents what goes directly to the models, such as information from the latent space of the VAEs and/or after PCA.

We trained the CNNs, MLPs, the predictor part of the VAEs, and the PCA models by the strawberry observations with connected measurements, which are 178 out of 304 Brix measurements. We trained the encoder and decoder parts of the VAEs by all the segmentation outputs of the *Mask R-CNN* model. Hence, this dataset includes images that were taken over the life cycles and of more strawberries. The *image-only experiment* and the *image-with-env experiment* applied the same encoders.

We designed the *env-only experiment* to analyze the relationship between the environment data and the Brix. We used rolling averages of the environment data over different periods. Since the environment data does not include specific information about individual strawberries, we took all of the 304 Brix measurements into account and grouped them by each harvest. They are called the “aggregated Brix”. The reliability of the aggregated Brix could also be better ensured by introducing more sample measurements. We not only trained machine learning models to predict the value expectation, but also the standard deviation (std.) and the percentiles from 10% to 90% (with intervals of 10%). The representations of the Brix distribution were considered in supporting further experiments of individual Brix prediction.

Since the amount of data points was reduced to the same as the days of harvests after the aggregation, the volume of the dataset became too small to support the training of deep neural networks. Hence, we applied linear regression (LR), support vector regression (SVR), and kernelized ridge regression (KRR) models. In addition, leave-one-out experiments were considered to enlarge the training sets of the *env-only experiment*. That means we split only one data point as the validation set in each experiment run, instead of proportionally splitting. Under this setting, we ensured all the data was used once in performance validation so that we could get a predicted value at every data point. The performance of individual Brix prediction in the *env-only experiment* is discussed based on the results from the leave-one-out experiments, by considering the predicted value expectation as the Brix predictions of all harvests on the same day.

## 4

In the *image-with-env experiment*, we stacked the features of images and the environment data according to the object strawberries to train models. By the encoder parts of the VAEs and the PCAs fitting to the training set, we encoded the images to image features. We trained the models of the *image-with-Brix experiment* by the same image features but with the outputs from the *env-only experiment*—predictions of the mean, std., and percentiles, etc. We established four neural network architectures to fit the various size of features in both the *image-with-env experiment* and the *image-with-Brix experiment*, including three three-layer MLPs and one four-layer MLP.

We used the Keras library to build and train the CNNs, VAEs, and MLPs in the experiments. All model training used the Adam optimizer ( $\text{beta1}=0.9$ ,  $\text{beta2}=0.999$ ) and a learning rate of 0.0003. We considered random rotation, mirroring, and flipping to augment the image data. When training the VAE, we also considered random scaling up to  $\pm 10\%$ . We used the Scikit-Learn library to conduct PCA and to construct LR, SVR, and KRR models in the *env-only experiment*. The KRR used polynomial kernels of degrees up to 3 and penalty terms of 1 and 10. These are all state-of-the-art implementations in data analytics.

For all experiments except with specific definitions, we split the data into 7:1:2 for training: testing: performance validation. We ran each experiment 15 times with a fixed series of data splits. All the deep learning models were trained on a Geforce GTX 1080 GPU under a maximum of 300 epochs.

## 4.3 Experiment Results

This chapter describes our research findings in four steps: (i) the exploration of the dataset that we collected; (ii) our conceptual methodology of designing the experiments; (iii) the model performance of each series of experiments respectively; (iv) two influencing feature selections: whether to use the plant load data or not and which image encoder to choose. The last section gives a comparison among the experiment series and states our suggestions for developing a reliable Brix prediction model.

### 4.3.1 An integrated dataset of growth and harvest quality

In order to predict Brix from non-destructive in-field data, we collected observations of the fruits and related environmental records in a greenhouse. The observations were in the form of images, and the environmental records are time-series and single-value mea-

surements. All relevant data were linked with the observations of individual fruits. As such, we could implement machine learning techniques to discover the mapping from the collected data to the Brix values.

From April 2021 to November 2021, we recorded the growth of strawberries by 13,400 images from three RGB cameras and collected environmental records during this period. We measured the Brix of 304 ready-to-harvest strawberries, which were selected from 28 harvests in 22 weeks. The overall statistics of the measurement data set are shown in Figure 4.2. According to the box plots and the line plot, the Brix at each harvest usually has a median value lower than the mean, implying that using the averaged sample measurements to estimate the Brix of every fruit has a higher probability to overestimate.

The environmental records during the data collection period were archived hourly and were grouped by rolling averaging over periods. As a preliminary analysis, we computed the correlations of the environmental data under different averaging periods and the aggregated Brix values of each harvest. The results indicate a strong correlation between temperatures (measured on the leaves, plants, and in the air), radiation levels, watering, and cyclic lighting strengths with the mean Brix of each harvest. The correlations of the Brix with humidity and CO<sub>2</sub> density are weaker. Details of the correlation analysis can be referred to Figure A11 in the appendix.

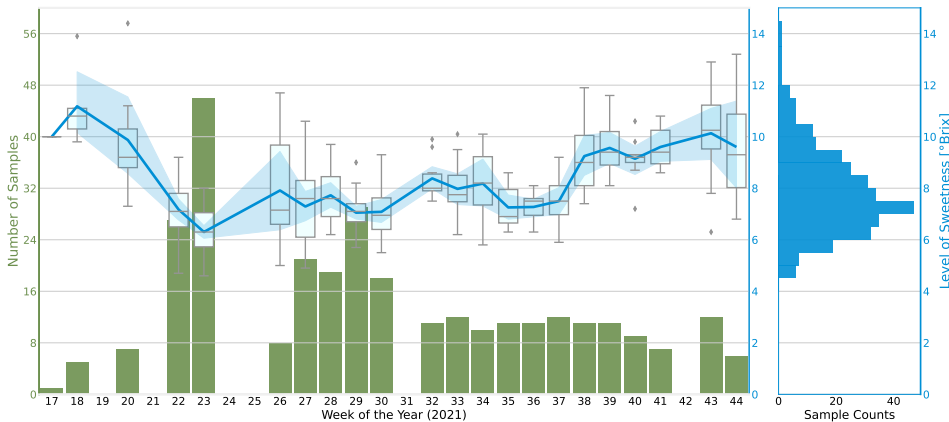


Figure 4.2: Statistics of the Brix measurements, grouped by harvests per week. On the left, the x-axis indicates the calendar week number of the harvests. The green y-axis presents the number of tested samples. The blue line and its contour indicate the averaged Brix value and the standard deviation (std.) of the measurements of the week respectively. The box plots illustrate the distribution of the measurement for the week. On the right, the histogram gives an overview of the distribution of all Brix measurements in 2021.

### 4.3.2 Conceptual experiment design

We designed four series of experiments to study the effectiveness of using these data, shown in Figure 4.3: we first analyzed whether the images (Section 4.3.3) or the environment data (Section 4.3.4) could work alone in Brix prediction, and then we considered two ways of data fusion (Section 4.3.5).

In *image-only experiment*, the Brix prediction model was trained solely by the images of strawberries. We considered both supervised learning (SL) and semi-supervised (SSL) in training the models in this experiment series. A challenge in this experiment was that the inclusion of non-relevant pixel data lowered the learning process and even reduced the prediction accuracy. To reduce this effect, some of the models were accompanied by additional regularization, such as conducting principal component analysis (PCA) on the training dataset and using the principal components as the features for learning.

We considered environmental records and/or plant loads as the input in *env-only experiment*. Together we call these the environment data. In the primary step, we conducted correlation analysis to classify the importance of each sort of attribute and to define sets of features. Since the environment data cannot express information about individual strawberries, we trained regression models to predict the expectation and the distribution of Brix value aggregations of each harvest.

4

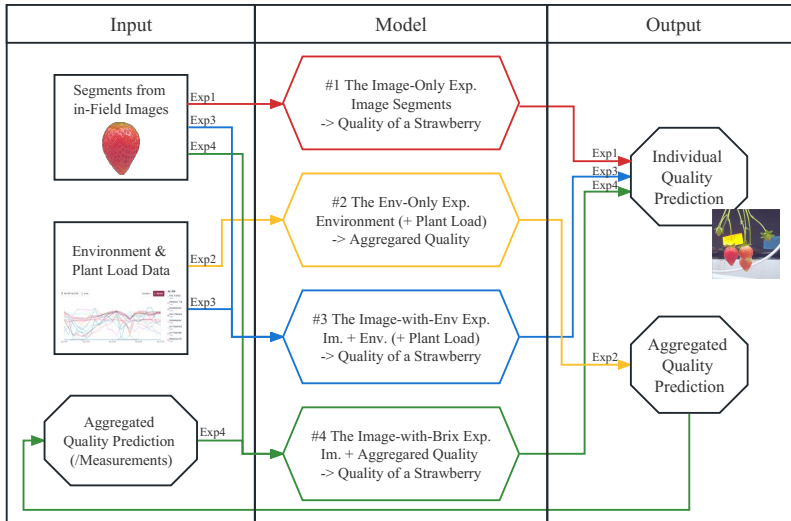


Figure 4.3: The methodology of the four experiment series in this research. They are described by the data flow, consisting of the input attributes, the output objects, and the models to map the corresponding inputs and outputs. The line colors and the short notes indicate different experiment series: red represents *image-only experiment* ("Exp1"), yellow is for *env-only experiment* ("Exp2"), blue is for *image-with-env experiment* ("Exp3"), and green is for *image-with-Brix experiment* ("Exp4"). All the models are evaluated by comparing the outputs with the ground truth.

We established *image-with-env experiment* and *image-with-Brix experiment* respectively as two ways of integrating the image data and environmental records in training. We encoded the image of each strawberry to comprise the image features. These features were combined directly with the environmental records to train the neural networks in *image-with-env experiment*. We considered the image features and the aggregated Brix predictions from *env-only experiment* as the inputs in *image-with-Brix experiment*. The setup was chosen based on two assumptions: (i) the predictions from *env-only experiment* are good indications of the overall quality of harvests; (ii) compared to predicting the absolute

Brix, the appearance information might be more helpful in terms of estimating the relative position out of value distribution of Brix.

We set up two baselines to evaluate the experiment outcomes. First, we used the average value of all the Brix measurements as the expectation of the *Favori* species. It represents the empirical Brix value that members of the soft fruit supply chain usually believe, so it is named the *empirical baseline*. It is the baseline of this Brix prediction study. Second, we considered the average Brix of each harvest as the expected value. As it represents the traditional way of sugariness assessment, which is anticipated by sample measurements, it is called the *conventional baseline*. According to the experiment setup, the *conventional baseline* is essentially the optimal situation of models from *env-only experiment*.

We used root mean squared error (RMSE) and mean absolute error (MAE) to represent the model accuracy. The RMSE is regarded as the main indicator of model performance. It gives increasingly more punishments if the predicted value is further from the ground truth. After running the experiments over different dataset splits, we used the standard deviation of the RMSEs (RMSE-std.) to indicate the robustness of model performances. The coefficient of determination (also called the R2 score) is considered a quantitative assessment of the level of model fitting. It is the proportion of the variation in the dependent variable, i.e. the individual or the aggregated Brix in this case, that is predictable from the input data. Higher R2 scores indicate better correlations between the inputs and outputs in the mapping.

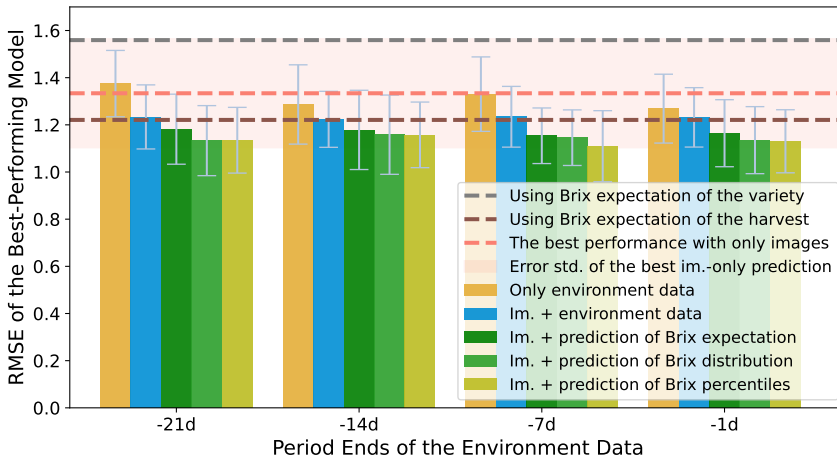


Figure 4.4: Performance comparison of Brix prediction accuracies among the four experiment sets, using RMSE as an indicator. The error bars indicate the standard deviation of RMSEs (RMSE-std) across different splits of validation sets. The models are grouped by the ending point of the periods of the environmental records. The y-axis shows the minimum RMSE of models from the same group. The colors indicate the input attributes of the experiment sets. The best performance of models using only image data is presented by a horizontal line. The contour around it indicates the corresponding RMSE-std. The horizontal line in gray and brown indicates the two benchmarks that are mentioned in the methodology section.

### 4.3.3 Brix prediction models require more than images.

By *image-only experiment*, we inspect the feasibility to train a Brix predictor with only images. We trained CNNs from scratch, with transfer learning (TL), and with semi-supervised learning (SSL) methods. The best-performing model of the entire experiment series has an averaged RMSE of ca. 1.33 over different validation splits.

As the horizontal lines in Figure 4.4 indicate, the selected model outperforms *empirical baseline*, while it is slightly worse than *conventional baseline*. It is implied that the appearances of strawberries provide hints of the Brix to a limited extent, whereas the time of harvest has more predictive power. We hence conducted further experiments to unravel the other attributes for Brix prediction.

Among the experiment results, we noticed that the involvement of feature dimensionality reduction facilitates the model performance. A possible mechanism would be that a large proportion of overlapping features were condensed in the latent space of VAEs or the orthonormal bases of PCA [95]. As the pixel data from a fine image is likely to correlate with each other, PCA is practical to de-correlate the data and facilitate model training. Meanwhile, the model fitting might also be regularized with the help of PCA, particularly when the model was trained with a small data set in our situation [66, 89]. These findings also motivated us to encode the images in the data fusion steps of further experiments.

### 4.3.4 Models reveal dependencies on environment data.

#### Performance in predicting aggregated values

In *env-only experiment*, we trained LR, SVR, and KRR models to assess how well the collective Brix value can be predicted with only the environment data. When aggregating the data points, overfitting was an indispensable issue. Particularly, when the data are very few whilst the inputs have a large dimension. To assess the level of model-fitting, we calculated the R2 score of models using different subsets of features, hyper-parameters, and train-test splits to predict the representations of value aggregations on the testing data set. When we grouped the scores by the algorithms of models to evaluate the level of model determination, we found more than half of the LR models have a negative R2 score, which indicates that simple linear models cannot fit this mapping. With a stronger regularizer, or with higher outlier flexibility, the R2 scores of KRR ( $\alpha=10$ ) and SVR models are more condensed to 0.5-0.6. The generally higher R2 scores also indicate they are more practical models in tackling this circumstance.

#### Performance in predicting individual values

To make the results comparable, the predictions of the averaged Brix were regarded as the estimation of all the strawberry measurements at each harvest. The RMSEs were hence calculated on the same validation splits as the other experiment sets take. Figure 4.4 compares the effectiveness of using various periods of environment data with other experiments, of which the time spans are grouped by the ending time.

As the bars in Figure 4.4 demonstrate, when models use features from the periods closer to the harvest time, they obtain lower and less diverged RMSEs in general. The RMSE-std of the models in *env-only experiment* is lower than the best-performing model from *image-only experiment*. The result argues that even using only the environment



data in Brix prediction could train more reliable and stable models. Hence, it is strongly suggested to involve the environment data in training further comprehensive models.

### 4.3.5 Images enable individual prediction with env-data.

Results from *env-only experiment* indicate that we need specific information to distinguish fruit-to-fruit differences from each harvest. Since the environment data are all point measurements, we encoded the images into 200, 50, 10, and 5 features by four combinations of VAEs and PCA respectively to fit the dimension differences between the two types of data. The *image-with-env experiment* and *image-with-Brix experiment* introduce two ways of fusing the image feature and environment data.

#### Combining image features with direct environmental information

The *image-with-env experiment* straightforwardly combined the two types of data to train the MLPs for the individual Brix prediction. Unsurprisingly, the lowest RMSEs from all the groups outperformed the best models from *image-only experiment* and *env-only experiment*, as is illustrated in Figure 4.4.

As is shown, the performance difference caused by the collection time span of environment data was remarkably reduced in this experiment. A possible reason would be that the MLPs also learned the trend of changes within the time-series data – such that the performance did not reduce as much as in *env-only experiment*. Meanwhile, the non-linearity and regularization performed by the neural network also ensured the robustness of the model performances.

#### Combining image features with predicted distribution of a harvest

The fourth experiment, *image-with-Brix experiment*, allows us to explore another way of integrating the knowledge from the two sorts of data: to use the image features to predict the relative quality within the distribution of Brix values. We used the predictions of Brix aggregations<sup>1</sup> from the leave-one-out experiments from *env-only experiment*. Among all the experiment series, the models from *image-with-Brix experiment* resulted in the lowest RMSEs, as illustrated in Figure 4.4. Among the different features of the aggregated Brix, models that were trained by Brix percentiles slightly outperform the models that assumed a Gaussian-distribution fit, i.e. using the mean and std. as inputs.

### 4.3.6 Plant load further facilitate prediction performance.

As is illustrated in Figure 4.5, introducing the plant load as part of the input attributes has a positive effect on the model performances, which is more outstanding on the models from *env-only experiment*. In *image-with-env experiment*, the upper limit of model accuracy was slightly improved. But more importantly, there were notable decreases in the std. of RMSEs over different data splits. Both changes were limited in *image-with-Brix experiment*. In all, we suggest that plant load is a crucial feature when the raw environmental information comprises the input data.

Moreover, since our plant load data was averaged over different branches of strawberries, they did not directly reflect the division of nutrition on the camera-monitored plants

<sup>1</sup>To limit the variables, we took only results from the KRR model with  $\alpha=10$  and polynomial degree=3.



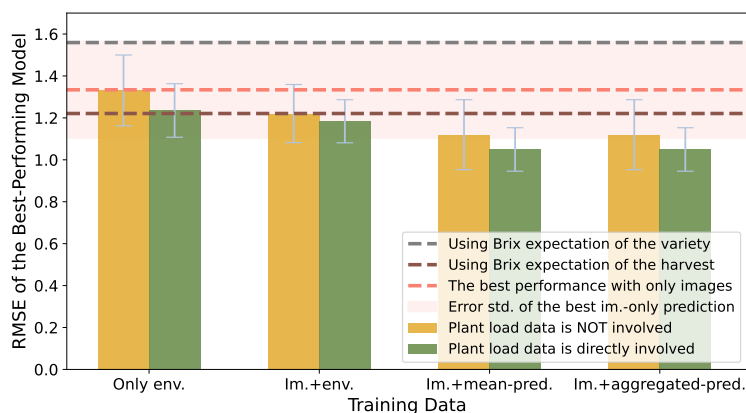


Figure 4.5: Performance comparison of Brix prediction using different attributes of environmental information, using RMSE as an accuracy indicator. The colors indicate the involvement of the plant load data. The y-values indicate the minimum RMSEs of models from the same group.

as the literature suggests. Hence, we suppose that the data could reveal the general influence of the growing environment on strawberries in this greenhouse compartment in an indirect and deferred way.

#### 4.3.7 Image encoders notably impact model performances.

The best-performing models of each family are considered in the previous result discussions. However, the number of image features also influenced the model accuracy. The information from different latent spaces is illustrated in Figure 4.6. Figure 4.7 discusses the effects when the image features are utilized with different representations of environment data.

When we used only the images in the prediction, it is still important to keep as many features as possible. Referring to the illustrations in Figure 4.6, it is indicated that considering the texture and the shape of strawberries could have a positive influence on the intrinsic quality representation. When using image features together with the raw environment data, we cannot see much difference in the best performances. Nevertheless, we observe an increase in the RMSEs when using larger dimensions of image features with the aggregated Brix. Overall, it is suggested that similar dimensions of image features and the other source of data could generally achieve better RMSEs.

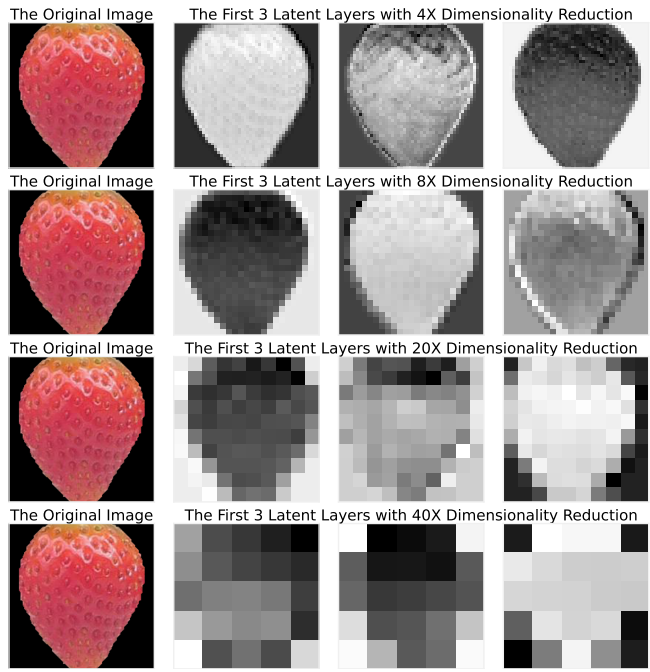


Figure 4.6: Examples of an image segment and its latent features from the four VAEs, plotted in a monologue style. The first column is the original image segment uniformed into a size of 200×200 pixels. The segment background is saved as black and transparent pixels. The level of dimensionality reduction from each encoder is shown on top of the latent space illustrations.



Figure 4.7: Performance comparison of Brix prediction using different image encoders, using RMSE as an accuracy indicator. The x-axis indicates the input attributes of the experiment sets. The colors indicate the dimensionality of the image features involved in the experiments. The y-values show the minimum RMSEs of all models from the same group.

Table 4.1: Detailed accuracy indicators of the best-performing models using different sets of input attributes. The models are ranked according to the "RMSE" column. the *empirical baseline* is calculated by using the Brix expectation of the strawberry variety as all the predicted values. the *conventional baseline* is calculated by taking the average Brix of each harvest as the individual predictions. The MAE and RMSE of all models and benchmarks are calculated by averaging over 15 random validation splits. The std. of the RMSE on each validation split is presented in the "RMSE-std" column.

Image Feature	Env. Data	Plant Load	Brix Agg.	MAE	RMSE	RMSE-std.
Included	In Agg. Pred.	In Agg. Pred.	Percentile	0.81	1.10	0.158
Included	In Agg. Pred.	In Agg. Pred.	Mean+std.	0.86	1.12	0.139
Included	In Agg. Pred.	In Agg. Pred.	Mean	0.90	1.15	0.118
Included	Included	Included	N/A	0.90	1.18	0.103
Included	Included	Not included	N/A	0.90	1.22	0.119
<i>The conventional baseline</i>				0.91	1.22	0.151
N/A	Included	Included	N/A	0.96	1.24	0.128
N/A	Included	Not included	N/A	1.00	1.27	0.146
Included	Included	Included	N/A	1.04	1.32	0.134
Included	Not included	Not included	N/A	1.00	1.33	0.189
<i>The empirical baseline</i>				1.21	1.56	0.312

### 4.4 Conclusion

In this chapter, we propose and evaluate a practical methodology for estimating the sugariness of individual strawberries, starting from planning the data collection setups. This approach uses affordable devices to collect relevant observations in the field and does not require harvesting or destroying the fruit. The experiment results demonstrate that it is feasible to anticipate the quality of strawberries when they are still growing. Such information could support the decision-making of harvesting and supply-chain strategies of greenhouse managers.

According to Figure 4.4 and Table 4.1, the models using image features with aggregated Brix information are the optimal choices among all the attribute combinations. The models could reduce the RMSE by up to 28.8% and 18.9% from the *empirical baseline* and the *conventional baseline* respectively. Compared to the image data, the environmental information has shown to be more relevant for the models to learn from, yet they lack the capability to tell fruit-to-fruit variances. Compared to using data from a sole source, a mixed-use of both could lead to an accuracy improvement of 10.0% and 6.2%, respectively.

Compared to other research in the field, we included multiple types of data to build machine-learning models. Our models show competitive performances in the sweetness prediction of strawberries – e.g. RMSE 1.2 from Sun et al. [238], RMSE 1.18 from Amoriello, Ciccoritti, and Ferrante [10], MSE 0.95 from Cho et al. [55] – while using in-field data collected more easily-acquired devices. On top of that, the dataset that we collected for pursuing this research is also useful for more research in this field.

In the aforementioned experiments, we performed all the procedures step-by-step, yet we see potential for greater model integration. Nevertheless, as state-of-the-art computer vision technologies allow detection models to be faster and more portable, expanding the capability of real-time assessments of fruit quality could also be an interesting topic.

The research primarily studies in-field and non-destructive data that are worth to be considering in training Brix prediction models. The images, which the prediction models were trained with, are essentially a subset of the time-lapse image dataset. With the entire dataset, further research is suggested to include temporal information for refining the quality prediction models. It is also an interesting topic to explore the practicability of using earlier images in forecasting future Brix values.

Our results suggest that environmental information plays a vital role in training a reliable model. Particularly, the environmental information from up to fourteen days before the harvest is crucial to ensure the model's accuracy. Nevertheless, we did not discuss the detailed influence of specific sources of climate data on our model accuracies. It is therefore recommended to conduct subsequent studies on the effectiveness of learning with different environmental factors.



## 5

## Data-Driven Precision Harvest Planning via Visual Growth Forecasting

## 5

This chapter demonstrates a framework to support fruit-specific decision-making for harvests, using strawberries as a use case. This framework addresses operational-level optimization, creating precise harvest plans to meet demand established by longer-term strategies. The optimization process starts with the methods introduced primarily in Section 3.4.5, which quantifies ripeness throughout the growth of a strawberry. For the fruits that hold sufficient historical observations from growth monitoring, we forecast their future growth using ripeness as the major metric and assign them to harvests at different dates in the future, facilitating informed decision-making for optimal yield.

This chapter is under the reviewing process of: Junhan Wen, T. Abeel, M. de Weerd. "Data-Driven Precision Harvest Planning via Visual Growth Forecasting." Applied Artificial Intelligence.

*While Artificial Intelligence (AI) techniques have been widely applied to problem-solving, their true transformative power lies in enhancing people's practice. For instance, in greenhouse agriculture, data analytics can enhance crop monitoring and resource planning; computer vision can support detecting crops and their conditions; and machine learning methods can foster tasks such as quality assessment and yield prediction. However, these technologies are often deployed only in isolated stages of cultivation, highlighting the absence of an integrated pipeline that connects them to what is desired in real world practices. To bridge this gap, we propose CROPi, an end-to-end framework that leverages Computer vision for Resource Optimization and Precision harvest using Image data.*

*CROPi is developed with four core functionalities: i) we propose and justify using hue as an objective redefinition of fruit ripeness; ii) we parameterize the ripeness development over time by analyzing growth monitoring images; iii) we develop predictive models to forecast the growth of new fruits; and iv) we optimize fruit-level harvest plans by aligning predicted growths with sequential demands. We demonstrate this approach using strawberries as the use case. By integrating insights from data mining and machine learning into harvest planning of crops, CROPi lays the foundation for next-generation precision agriculture, ensuring efficiency, sustainability, and improved yield management.*

## 5

## 5.1 Introduction

Effective and efficient planning in real-world scenarios often calls for a combination of diverse, interdisciplinary techniques to extract meaningful structure and insights from data in the wild. Such data are typically multimodal, noisy, and sometimes scarce, demanding the methodical integration of complementary methods to bridge the gap between heterogeneous inputs and abstract, behavior-level targets. For example, when optimization algorithms are used to guide future plans, their utility depends heavily on accurate representations of the current state. This, in turn, necessitates the use of artificial intelligence (AI) to process unstructured inputs into reliable, real-world assessments and forecasts. However, many existing technologies focus narrowly on isolated subtasks, resulting in a lack of integrated, end-to-end workflows. Addressing this gap requires a paradigm that unifies data-driven methods to orchestrate insights from diverse sources in a coherent and task-oriented manner.

A typical field that is currently being reshaped by information and data science is agriculture, leading to the term “precision agriculture”, where in-field data is converted into actionable crop-management decisions [64]. Machine learning models can be trained on such data, which are primarily images, to support crop counting, disease and pest detection, yield prediction, and fruit-level cultivation by humans or robots [20, 76, 172, 197, 212]. Yet most solutions still tackle narrow, stand-alone tasks – such as training a ripeness classifier that merely mimics humans’ judgment. Developed in isolation, these models are hard to weave into the broader pipelines that real-world decision-making demands. What growers need instead is integrated, interdisciplinary data analytics that delivers objective, actionable insights. Hence, we introduce CROPi, an interdisciplinary, end-to-end pipeline that combines machine learning, computer vision, and algorithmic components into a seamless pipeline for a realistic challenge: using accessible infield data to determine optimal harvesting strategies, which can assist both immediate and future decisions.

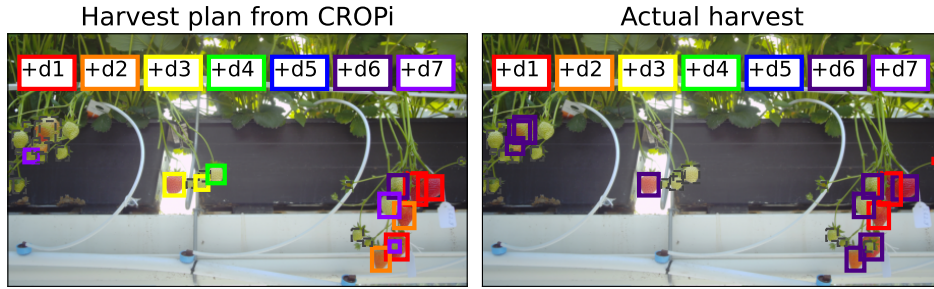


Figure 5.1: The left image shows a plan from CROPi under a daily demand of 3 fruits; the right shows the practice of the greenhouse in the same period, where “+d\*” means harvesting \* days from now.

The primary motivation of CROPi stems from growers’ interest in engaging in “contract farming” – a marketing arrangement in which they sell most of their agricultural output to downstream buyers, particularly wholesalers with sufficient scale to influence bargaining power. Contract farming is increasingly popular because its pre-arranged terms for supply quantity, quality, delivery schedules, and pricing provide growers with a stable market, especially for perishable products like soft fruits [116, 194]. However, many growers involved in these arrangements are of a small scale and operate independently, which can place them at a disadvantage position when negotiating with larger buyers who hold greater market power [24, 116]. For instance, the consequences of breaching such contracts are severe, and risks related to harvest uncertainties lie mainly with the growers [24, 194], although this cannot be perfectly controlled. Consequently, this forces the growers to continuously adjust decisions in their daily harvest practices, e.g. re-deciding the minimum maturity of fruits to be harvested, to meet the contracted demands without exceeding them by much. As of now, with limited tools to quantify maturity and model growth, decisions largely depend on the growers’ expertise. Managing these complexities requires more advanced, data-driven approaches that can account for both analyzing growth dynamics and optimizing harvest timing.

To this end, we develop CROPi: an end-to-end pipeline that leverages temporal, in-field data to optimize sequential harvest decisions. Combining computer vision, machine learning, and optimization, CROPi supports precision agriculture by aligning growth monitoring with downstream demand: in the first place, we propose using hue values as an objective alternative to eliminate inconsistencies in subjective quality assessment. This refined ripeness definition allows us to quantify the entire growth process, which can be modeled using standard growth functions [210]. Subsequently, forecasting models trained on cultivation records then predict future growth from early-stage data and promote proactive decision-making for the greenhouses. Finally, these growth predictions are combined with future demand to optimize sequential harvest decisions together.

To the best of our knowledge, no existing approach integrates biological development analysis with future demands to arrive at optimizing the harvest planning. CROPi bridges this gap by mining key information from temporal monitoring data (which are mostly



images), using this for forecasting daily production, and matching this to daily demands. Specifically, it converts time-lapse images into fruit-level time series that support predicting the maturity levels during growth, and disaggregates higher-level harvest demands (e.g., in kg) into fruit-specific targets per day, so as to enable both sides of the optimization problem to be expressed at the fruit level. Through this alignment, CROPi enables integrated, fruit-level optimization using data accessible in the field, which further paves the way for the broader adoption of more autonomous, data-driven farming techniques.

In this chapter, we start with a detailed problem definition as in Section 5.3. We introduce the conceptual workflow of CROPi by the specified methodologies behind each functionality in Section 5.4. The workflow is illustrated using realistic strawberry growth records from greenhouse practices in Section 5.5. Strawberries are chosen as a use case due to their typical growth pattern, which is indicated by hue, and the availability of open-source image datasets that provide individual-fruit level monitoring [272].

Overall, the contributions of this chapter lay in:

- a comprehensive, end-to-end pipeline that enhances precision agriculture by leveraging multi-modal greenhouse monitoring data;
- a vision-based, objective method for assessing fruit maturity in the field and enabling parametric growth modeling;
- an effective and efficient approach to optimize sequential harvest decisions based on growth and demands.

## 5

## 5.2 Background and Related Work

### 5.2.1 Current practice with vision data

The development stages of many crops are defined by the change in appearance during the biological development of the plants. For fruits that show an evident color change among stages, color cards are designed to improve the consistency for people to assess the maturity level [16, 143]. Hence, image analysis and machine learning are widely adapted to predict ripeness [172]. The intuitive development of models, predicting human labels from images, can also be easily plugged with developed object detection frameworks to perform on scale. Nonetheless, in order to lift their efficiency, workers may rely more on their experience than the color cards in practice. While most existing works focused on classifying the maturity stages that are assessed by people, they essentially introduce subjectivity to the objective assessments. Thus, in this chapter, we suggest redefining ripeness labeling by quantitative color analysis to achieve true objectiveness in the modeling. We believe this leads to more consistent results and better defines a continuous space of ripeness rather than manual categorization with vague and potentially overlapping boundaries.

To use color as the primary characteristic for defining ripeness, it is crucial to minimize illumination variations caused by factors such as the time of image capture and changes in viewing angles [218]. Preliminary research on identifying fruit ripeness using image data was usually taken out with laboratory settings to limit the influences [120, 144]. In fact, computer vision techniques such as color correction algorithms [92, 94] and generative algorithms [42, 150] are practical for conducting robust and adaptive corrections. We believe these promising candidates can facilitate the use of infield data, which lead to higher applicability of the entire machine learning framework.

## 5.2.2 Formulation of growth functions

In agricultural research, parametric functions have been used to model crop growth, including overall development, future, and even seasonal production (we use the term “growth function” from here on). The sigmoid function, or the more generalized form, Richard’s function [210], is a widely-used mathematical model to parameterize the growth for many crops [90, 118]. In research where the growth models were considered to optimize harvests, researchers took fixed coefficients from statistics of species [104]. However, when the other influencing factors, such as the growing environment and cultivation treatments, are omitted, the growth models are not precise enough to provide realistic predictions applicable to actual harvests [111].

A major reason for this limitation is the lack of comprehensive analyses that account for variations in time, location, and treatment conditions. One promising way to address this gap is through the mining of crop growth monitoring data, which, as previously discussed, provides a directly accessible source of information for capturing these variations. In this context, computer vision techniques play a key role by enabling objective analysis and helping extract consistent, generalizable insights from the data.

## 5.2.3 Prediction of future growth

Looking from the dimension of time, growth functions are time-ordered data, shaped by species-specific patterns but influenced by environmental conditions. Forecasting growth, then, becomes a task of completing the function based on partial observations, i.e., predicting future development from what has been observed so far. This forecasting can also be framed as a next-token prediction task, where the model continues the observed sequence by mimicking its underlying process. This approach, common in natural language processing [3], has also proven effective in traditional time-series forecasting [188, 296].

Alternatively, since the growths tend to follow a similar pattern, k-nearest neighbor (k-NN) and clustering methods can help advance analysis by capturing temporal dependencies and complex trends through distance-based comparisons [230]. Besides traditional measures like Euclidean distance, specialized metrics such as autocorrelation and Dynamic Time Warping (DTW) [217] could be more practical to be considered. When the data contains noise or perturbations, distances like softDTW [62] and LB Keogh lower bound [131] are preferred, as they offer greater robustness by accommodating temporal mis-alignments and outliers. With these metrics, the aforementioned machine learning methods could work more effectively.

## 5.2.4 Harvest and supply optimization

Existing research on agri-food supply optimization has typically addressed broader, more general problems, on which they often employed linear programming or dynamic programming models to incorporate uncertainties and constraints [7, 243]. The implementation of such methods tailored the solutions to focus on decision-making at the level of management units or batches [46, 97]. However, this approach remains distant from the practical realities faced by harvest robots. Alternatively, in solutions for other practices, when framed the optimization as an assignment problem and subjected to proper feasibility checks [40, 148], the Hungarian algorithm can achieve optimized planning within polynomial time [138, 142].

## 5.3 Problem Formulation

### 5.3.1 Context and Definition

In this chapter, we investigate a real-world use case that integrates analytics of fruit (growth) development to the optimization of harvest decisions. For non-climacteric fruits like berries, which do not keep ripening after picking, their maturity at harvest directly determines selling quality. Therefore, CROP1 primarily focuses on optimizing the timing of harvest by aligning fruit growth stages with delivery requirements.

The problem involves two fundamental sources: on one side, cameras in greenhouses monitor fruit growth through videos or time-lapse images, enabling us to reconstruct past development and predict future ripeness by associating changes with frame timestamps. On the other side, delivery contracts provide known requirements, specifying quantities and deadlines in advance. Hence, we could model the problem as an assignment optimization: allocating harvest times to fruits to best meet quality and delivery targets.

### 5.3.2 Formulation

Overall, the problem is formally defined as:

- **Given:**

- The growth monitoring of fruits  $j \in \{1, 2, \dots, N\}$ , in the form of video frames or image series.
- The demand from downstream  $D_t, t \in \{1, 2, \dots, T\}$ , which are sequential and are fixed from contracts.
- Other related records, e.g. weather in the past.

- **Find:**

- The optimal plan, defined as a set of binary actions  $x_{jt}$  (“harvest or not”) over time  $t$  for each fruit  $j$ .

- **Objective:**

- Maximize the profitability of selling by planning  $x_{jt}$ :

$$\max(\sum_{t=1}^T \sum_{j=1}^N x_{jt} \cdot V_{jt}) \quad (5.1)$$

- where the profit  $V_{jt}$  is assessed by scoring the growth of fruit  $j$  at the proposed harvest time  $t$ .

- **Constraints**

- As a natural constraint, one action per fruit, i.e.  $\sum_{t=1}^T x_{jt} \leq 1, \quad x_{jt} \in [0, 1]$  for  $j \in \{1, 2, \dots, N\}$ .
- Each growth is modeled as an independent time-dependent function  $G_j(t)$ , i.e. harvesting  $j$  will not significantly impact others'  $G_{-(j)}(t)$  of the plant.

### 5.3.3 Data source

In the demonstration of this chapter, we primarily utilized public image datasets to evaluate the approach under a spectrum of demands. Our inputs cover the following categories:

**Greenhouse monitoring:** “The Growing Strawberries” Dataset (GSD) [272] collected on-site time-lapse images during cultivations in two years. Example raw images are depicted in the first row of Figure 5.7. Its annotations for a multiple object tracking (MOT) task identify the growth of individual strawberries, from which we can curate the datasets for growth monitoring and forecasts in the following demonstration.

In addition, we used the timestamps of the monitoring images to align the cultivation record with close-site weather data [109], so as to provide more input features for the prediction task.

**Future demands:** Due to the lack of relevant public data, we designed a set of typical comparative scenarios to simulate the demands required for our analysis. These scenarios represented cases in which demand could be higher than, close to, or lower than projected future yields. The synthetic data was crafted to reflect realistic conditions and generate meaningful insights for the study.

**Fruit maturity assessment:** In addition, we discuss the outcomes of our proposed objective ripeness assessment in Section 5.5.1, and compare them with descriptive assessments or discrete scores assigned using typical datasets provided by Elhariri et al. [76] and Wen et al. [266] respectively.

## 5.4 Methodology

Figure 5.2 depicts the workflow of our proposed solution. The solution utilizes accessible on-site data from real-world greenhouses to optimize the sequential harvest decisions based on predictive growth models of fruits. so as to support the decision-making of how much and which fruits to harvest for the best possible match to the demand over time. The methodologies of the four main functionalities, as identified by the diamond shapes, are introduced in the following subsections, and demonstrated in Sections 5.5.1 to 5.5.4 respectively.

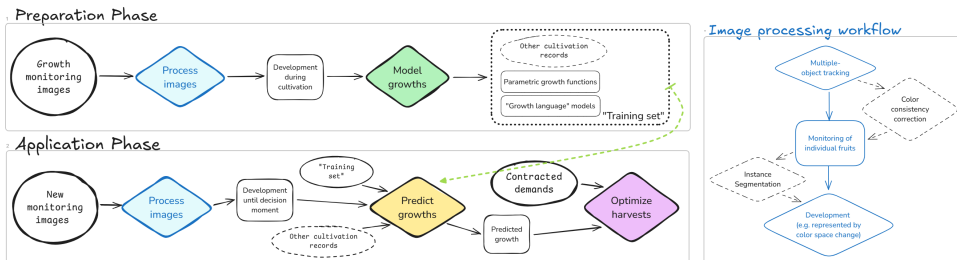


Figure 5.2: Pipeline of the CROPi framework, divided into two main sections: the Preparation Phase and the Application Phase. The core functional components, represented by the four color-coded diamond shapes, are benchmarked and discussed respectively in subsections of Sections 5.4 and 5.5. Specific steps of the first stage, “process images”, are provided within the additional section on the right.

### 5.4.1 Processing the images

The right subplot of Figure 5.2 illustrates the main and branched streamlines of the image processing workflow. Given videos or image series capturing crop growth, multiple object tracking (MOT) algorithms can identify individual strawberries and therefore tell the individual growth pattern of every fruit. To reduce noise in the further processing, segmentation models, or taking a subarea in the detected bounding box as an approximating alternative, can be applied on top of the MOT results. In the meantime, since many cultivation environments use only limited lighting control, using color consistency correction methods, such as the Gray World [94], intrinsic color decomposition [150], or style transfer models [42] can also be added to the main workflow.

We chose strawberries as an example because fruit color is a widely accepted indicator of maturity [119], making it a simple yet representative case for demonstrating CROPi. Traditionally, ripeness is assessed by comparing fruits to reference color cards or personal expertise. While computer vision offers more consistent and objective evaluations, current methods often rely on subjective human annotations to train models from scratch or fine-tuning from object detectors like YOLO or R-CNN. Since biological development is continuous and lacks clear stage boundaries, we propose using the fruit skin's hue to eliminate human error and ensure true objectivity. For strawberries, we used the  $A^*$  value from the CIELAB color space [221], which naturally reflects growth – from green on the negative axis to red on the positive axis. The  $A^*$  values collected over time form (*timestamp, color*) pairs that serve as the raw data for growth modeling.

### 5.4.2 Modeling the growths

We considered two ways of developing growth models: (1) a traditional method that fits development using a logistic growth function, referred to as the “parametric growth function”, and (2) a deep learning approach inspired by natural language processing (NLP), which models growth as a time-based sequence using an autoregressive next-token prediction (NTP) strategy. Given its roots in NLP techniques, we refer to the second approach as the “growth language” model.

#### Fit to parametric growth functions

Both Richard's curve and the logistic function are suitable for modeling the typical smooth-rapid-smooth growth pattern, as illustrated in Figure 5.7. Between them, the fewer coefficients for the logistic function, as shown in Equation (5.2), make it more robust and efficient for fitting noisy real-world data and thus is more applicable for our case [118]. Example growth functions are shown as dashed lines in Figure 5.6. Its alignment with natural biological growth trends also allows us to identify potential detection or tracking errors by comparing the fitted parametric curve to the raw  $A^*$  value trajectories. Such a comparison also highlights potential errors in GSD annotations, helping us refine the dataset before training the predictive models.

$$f(t) = \frac{A}{1 + e^{-B(t-t_0)}} + C \quad (5.2)$$

Note that since MOT is not the main task of CROPi, we simply took the revised annotations in the demonstration in Sections 5.5.2 to 5.5.4. For future applications on new

monitoring data, we can still employ real-time MOT algorithms to keep track of each new fruit's growth.

### Train growth language models

Beyond parametric function models, deep learning offers a more flexible way to capture growth patterns. Since input data lengths vary in realistic applications, NTP is more practical to make full use of available data. Additionally, with the rise of large language models (LLMs), transfer learning using open LLMs can improve efficiency, particularly with limited datasets like those in agriculture. Therefore, in developing CROPI, we explored both training NTP models from scratch and from fine-tuning open-source LLMs to learn growth patterns. In experiments in Section 5.5.3, we applied *nanoGPT* [129] to build our *growth language* models. Given the high computational demands of GPT models, we limited each model's training time to a maximum of 36 hours on an NVIDIA A40 GPU.

### 5.4.3 Predicting the growths

The task of growth prediction involves estimating the development over a fixed time range, based on the ripeness states observed so far (i.e., MOT data up to that moment) for the strawberries in a given frame, optionally incorporating relevant monitoring data such as weather. Using the GSD dataset [272], which includes data from both 2021 and 2022, we created two train-test splits to evaluate performance: (i) training on one camera's image series from 2021 and testing on another camera's images from the same year; and (ii) training on all 2021 images and testing on a single camera's series from 2022.

When using the *growth language* model, we treated each series as independent and added separating tokens at the end. However, a fixed forecast length can cause the model to predict a separating token prematurely, triggering a “new” series to be predicted without any observation. To address this, we replace all predictions after the first separating token with the last valid token before it, until the forecast period ends. This approach reflects the idea that maturity cannot decrease, and the extended stay in the mature stage can be handled later using the scoring function described in Section 5.4.4.

When fitting the *parametric growth function*, we first filtered by min. length, max. initial, and min. ending  $A^*$  to ensure the task uses relatively more complete growth mon-

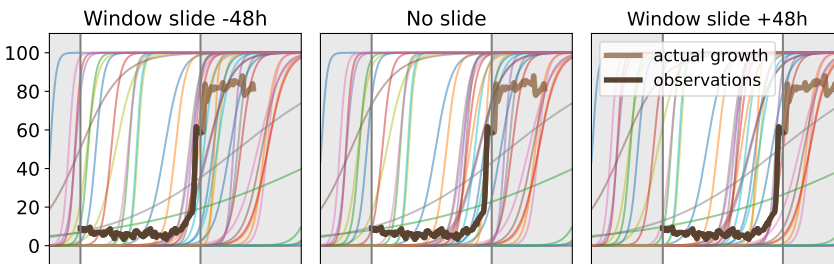


Figure 5.3: Demonstration of the sliding-window k-NN (s.k-NN) method. The observation is shown in dark brown, with lighter shades indicating the ground-truth future. Thinner lines are growth functions from the training set. For each observation, the  $k$  functions with the highest similarity and the corresponding window sliding pointers are noted to estimate  $T'_{j,0}$ .

itoring records throughout the seasons. These, together with relevant climate data from near-site sources, formed the “training set”. Since the training and testing sets cover different periods but share the same hourly resolution, we developed a two-stage algorithm to estimate growth function coefficients.

In the first stage, we identified growth functions from the training set whose inflection points ( $t_0$  in Equation (5.2)) fall within a narrow time window. The primary inflection point  $T'_0$  of the testing set was determined by aggregating votes from growth observations that have passed the lag phase [80], i.e. characterized by a clear transition from green to red, by comparing pattern similarities with historical growth observations or climate data in the training set. After narrowing down candidate growth functions based on  $T'_0$ , the second stage estimates potential inflection dates,  $T'_{j,0}$ , and other coefficients, i.e.  $A, B, C$  of Equation (5.2), using the similarity measures again, to identify the specific growth functions of fruit  $j$ . To assess these similarities, clustering, autocorrelation (autocorr.), and sliding-window k-NN (s.k-NN; as illustrated in Figure 5.3) were used for growth data, while clustering, 2D-autocorr., and k-NN were applied to 2D data of historical climate. S.k-NN with the same metric was used in the second stage if the first stage used a clustering method.

## 5

#### 5.4.4 Optimizing the harvest plans

The final task of CROP1 is to maximize the supply quality through planning the sequence of harvesting actions. We model this as an assignment optimization problem, where each fruit is assigned an optimal harvest time to ensure the best possible outcome. For non-climacteric fruits such as strawberries, which stop growing once harvested, the problem is equivalent to maximizing the overall quality at each fruit’s harvest moment. The reward of harvesting fruit  $j$  at time  $t$ , denoted as  $V_{jt}$ , can be assessed by the growth stage,  $G_j(t)$ . In Figure 5.4, we illustrate a few scoring options – such as binary, fuzzy logic, and fuzzy logic with timing penalties – which serve as examples but are by no means exhaustive.

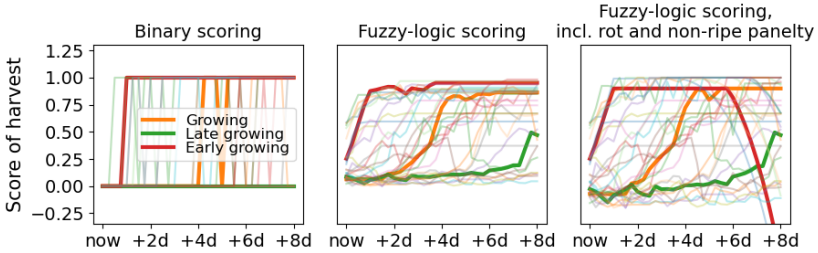


Figure 5.4: The scoring of harvests at different future moments, using predictions from our nanoGPT model. Each figure shows a different scoring method, using the same data. We highlighted the score change from three typical observation patterns, as noted in the legend.

We treated the harvesting decisions, denoted as a binary variable  $x_{jt}$ , as optimizing the assignment between fruit growths and harvest time slots. Hence, the fundamental form of the objective function, which is commonly referred to as the cost matrix in assignment problems, is defined as follows:

$$\begin{aligned} \max \quad & \sum_{t=1}^T \sum_{j=1}^N x_{jt} \cdot V_{jt} \\ \text{s.t.} \quad & \sum_{t=1}^T x_{jt} \leq 1, x_{jt} \in [0, 1] \text{ for } j \in \{1, 2, \dots, N\} \end{aligned} \quad (5.3)$$



Building on this formulation, we discretized the greenhouse-level demand based on visible plant area, resulting in integer-valued demands for each time slot, i.e.,  $D_t \in \mathbb{N}, \forall t \in \{1, 2, \dots, T\}$ . This setup enables the use of the Hungarian method to perform the optimization, which also guarantees that each element in the smaller set is matched to one in the larger set [138]. As a result, if supply exceeds demand, some fruits will not be scheduled for harvest; if demand is too high, even the greenest fruits may be assigned. To address these impractical scenarios, we introduce two additional options: an *oversupply allowance*  $D'_t$ , representing an alternative supply venue where the demand and/or price might be ( $\delta$ -)discounted; and a *postponed harvest* by placing a dummy demand variable  $D_\infty$ . The corresponding decisions are denoted as  $y_{jt}$  and  $z_j$ , respectively. With these options included, the objective function is updated accordingly:

$$\begin{aligned} \max \quad & \sum_{t=1}^T \sum_{j=1}^N (x_{jt} \cdot V_{jt} + y_{jt} \cdot \delta \cdot V'_{jt}) + \sum_{j=1}^N z_j \cdot V_\infty \\ \text{s.t.} \quad & \sum_{t=1}^T (x_{jt} + y_{jt}) + z_j = 1, \forall j \in \{1, 2, \dots, N\} \end{aligned} \quad (5.4)$$

Hence, when  $x_{jt} = 1$  or  $y_{jt} = 1$  in the solution, the optimal harvest time for that fruit  $j$  is the corresponding  $t$ . If  $z_j = 1$ , the fruit is excluded from this given harvesting period  $T$ .

Since no established metrics exist for evaluating fruit-level harvest activities, we assess the outcome plans by estimating the “reward”, i.e.  $V_j(t)$  as in Equation (5.4). In Section 5.4.4, we compare the plans optimized using predicted fruit growth with three benchmarks: (i) the actual harvest, derived from MOT annotations in *GSD*; (ii) a simple “harvest all mature fruit” policy, applied directly to the recorded growth data; and (iii) optimized plans using *parametric growth functions* fitted to those records (“complete simulation” as in Table 5.3). To ensure a fair comparison, we applied a fuzzy-logic scoring system that accounts for penalties, as later visualized in Figure 5.4, under the simulated parametric growth setting to calculate  $V_j(t)$ .

## 5.5 Experiments and Demonstrations

This section demonstrates the methodologies employed to execute the four principal functions of the solution pipeline, corresponding to the color-coded components in Figure 5.2.

### 5.5.1 Ripeness evaluation

To illustrate the necessity of using color as an objective measure to assess the ripeness, we first compared the exact color values from the fruit with the original maturity labels annotated by human experts. To accelerate, we used the averaged  $A^*$  value from the center crop of the detection bounding box (bbox). We compared these assessments with the original ripeness labels provided by the two datasets: *StrawberryDS* from Elhariri et al. [76] that uses six descriptions to categorize fruits from early growth until peak ripeness, and the quality assessment collection from Wen et al. [266] that uses a discrete grade from 1 – 10 to identify the maturity levels. Note that the latter dataset only consists of harvested strawberries due to other measuring purposes, resulting in the labels not being distributed over all values.

Figure 5.5 illustrates a trend where  $A^*$  values generally increase with higher ripeness levels. However, some overlap exists between distributions, especially when the categories are too fine. These overlaps likely stem from the subjectivity of human labeling



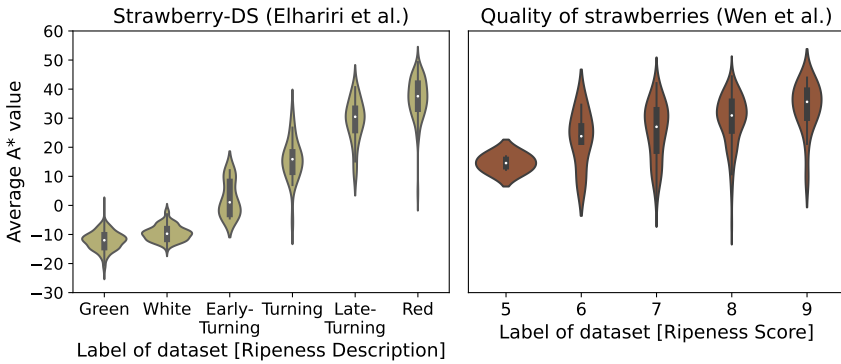


Figure 5.5: Distribution of  $A^*$  values of fruits in images under different ratings of ripeness. The left comparison uses the training set from Elhariri et al. [76]. The right plot uses the full dataset from Wen et al. [266]. The original annotation of each dataset is indicated on the x-axis of the corresponding figure.

## 5

or occlusion by other objects in view, with subjectivity being the primary factor. This is particularly significant because such ambiguity is even more evident in [266], where the classification boundaries are vague due to class refinement. Hence, the inherent differences between human and machine labeling efforts further underscore the need for objective metrics.

This experiment validates using hue (or  $A^*$  in this case specifically) as the primary reference rather than just a supplementary aid, to reduce subjectivity and improve consistency in ripeness evaluations. This approach not only supports more transparent and data-driven supply negotiations but also lays the foundation for building a quantitative, continuous growth model, as further demonstrated in the following functionalities of CROPi.

### 5.5.2 Growth function formulation

When we applied an objective ripeness evaluation as presented in Section 5.5.1 to process growth monitoring data, we could record the biological development as a time-based function. In Figure 5.6, we demonstrated this methodology using the growth monitoring images and MOT annotations of GSD [272].

As shown in Figure 5.6, the  $A^*$  value generally increases over the stages of biological development, though at varying rates. However, notable noise causes fluctuations, primarily due to changes in illumination conditions during the monitoring period. Among those, the most significant variations can be eliminated by filtering image data using a brightness threshold, as illustrated by the half-transparent lines in Figure 5.6.

To further mitigate the impacts of the varying illumination, we tested a few color correction methods on the image data, as demonstrated in Figure 5.7. The growth functions after selected pre-processing are shown as other colored lines in Figure 5.6. By comparing Figures 5.6 and 5.7, we observed that intrinsic color decomposition effectively minimized fluctuations, but at the cost of lowered image contrast and compressed hue value ranges. Generative models used for style transfer also improved color consistency from a perceptual standpoint; however, they sometimes introduced unpredictable sharp values, which adversely affected further processes. In contrast, although the Gray World algo-

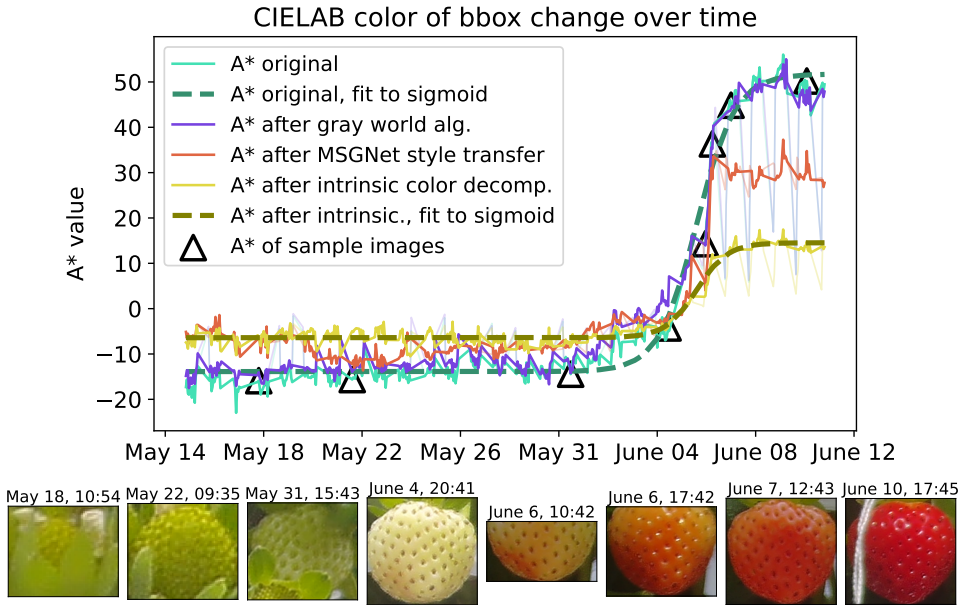


Figure 5.6: Example growth function (top) and observations (bottom) for a strawberry.  $A^*$  are shown for both the original and color-corrected images. Dashed lines represent sigmoid curves fitted to the  $A^*$  values. Darker lines indicate data from images with brightness above 50. Triangles mark the data points of the below observations.

rithm offered a more modest adjustment, its deterministic nature ensures that it behaves consistently and avoids the instabilities of generative approaches. Thus, we used its output in subsequent demonstrations.

### 5.5.3 Growth prediction

Given the available crop observations up to the decision-making point, we frame growth prediction as a completion task – estimating the full developmental trajectory from incomplete data. To this end, we explored two approaches: 1) training a *growth language* model to perform next-token prediction (NTP) in an autoregressive manner, and 2) fitting the *parametric growth function* from Equation (5.2) by seeking similar samples from observed patterns in the training set.

Tables 5.1 and 5.2 present the prediction results for the two train-test splits described in Section 5.4.3. Overall, NTP-based methods demonstrate stronger performance for this task. Notably, since most models trained from scratch did not fully converge within the budget computational time, the accuracy achieved by fine-tuning a pretrained GPT-2 model stands out more clearly. This advantage is particularly evident under cross-entropy loss, which favors sequence modeling tasks but is less suited for continuous regression.

In contrast, the pattern-matching approach via sliding-window k-NN (s.k-NN) achieved comparable accuracy in many cases but suffered from significantly higher computational costs. Within this family, matching based on growth patterns outperformed those relying on climate data similarity, while autocorrelation (autocorr.) was mostly ineffective.

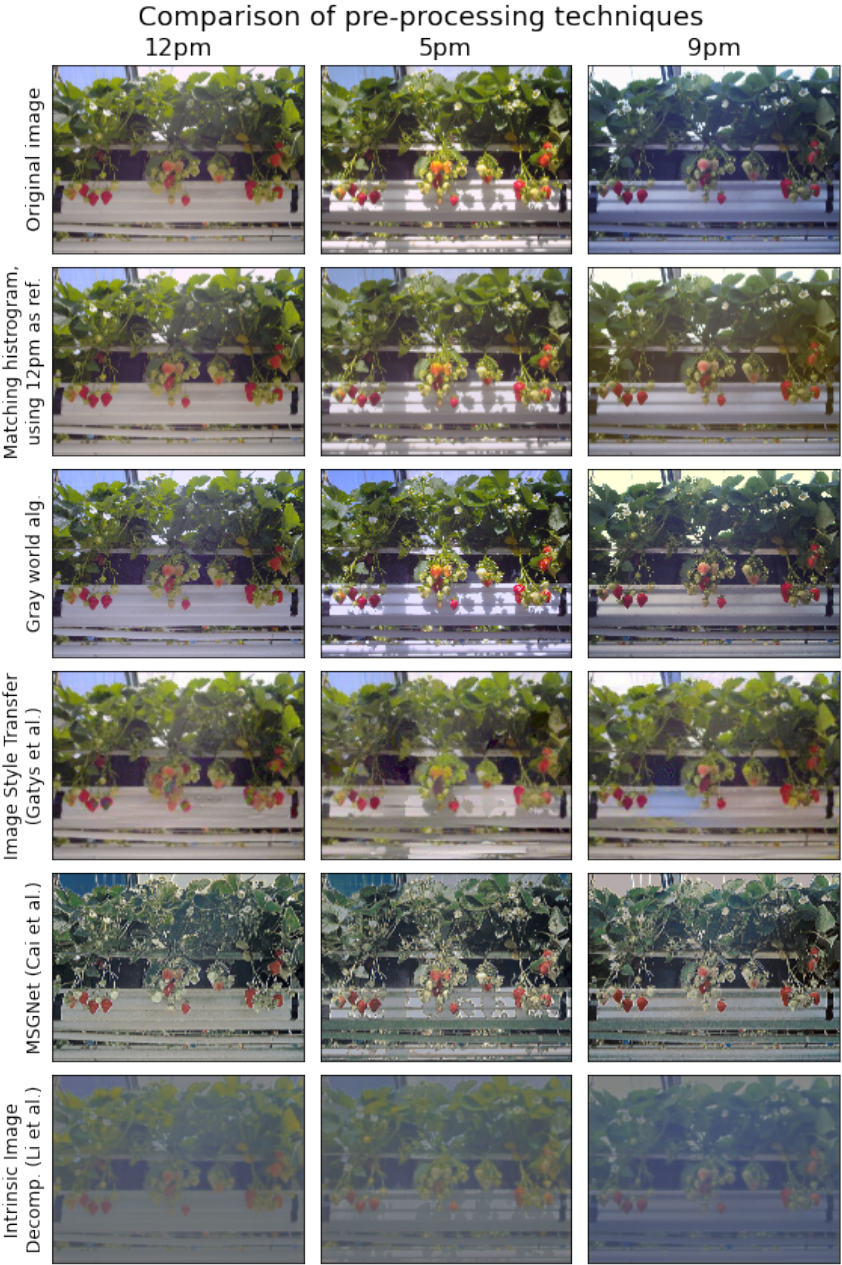


Figure 5.7: Illustration of color consistency corrections, using sample images collected at 12, 5, and 9 pm. The original images are at the top. The processing methods are noted on the left of the rows.

The reduced performance of autocorr. and other linear methods is mainly due to their tendency to select mismatched samples, particularly when the training and testing data come from different years. This often results in overly generalized, average-like predictions, as seen in the final two evaluations of Table 5.2.

Even in the testing phase, a trained model would take around 2 minutes to perform the NTP task. Hence, among the methods presented, s.k-NN with L2-norm offered a favorable balance between performance and runtime. In addition to this, as can be observed from Figure 5.4, outputs from NTP often retain fluctuations caused by illumination noise in the input data. To ensure clearer visualizations and more stable results for downstream analysis, we used predictions from *s.k-NN with L2* in the main demonstrations of Section 5.5.4. We also evaluated how different levels of prediction accuracy influenced optimization performance with more prediction results (with superscripted markers in Table 5.1).

Table 5.1: Comparison of selected growth prediction methods, measured on all visible strawberries on a random frame from 2021 as in Section 5.5.4. The table is split into four sections, based on the method and data sources. Performances are described by common time-series analysis metrics as in the header. Note that for two-stage methods, the same metric as in the second column was used in both. Superscripts link methods to their optimization performance in Table 5.3.

Method	Metric	L2 ↓	DTW ↓	AC ↑	Time [s]
PERFECT SIGMOID FIT		0.19	1.73	0.13	-
TRAINING A NEXT-TOKEN PREDICTION MODEL					
Full training	L2-norm	0.19	2.84	0.09	(max 36h)
Full training	CrossEntropy	0.33	3.68	0.12	(max 36h)
Finetune GPT-2 <sup>⚡</sup>	L2 <sup>⚡</sup>	<b>0.17</b>	2.30	0.10	(max 36h)
Finetune GPT-2	CrossEntropy	0.18	2.01	0.08	(max 36h)
CONSIDERING SHAPE SIMILARITY OF GROWTH FUNCTIONS					
autocorr.	autocorr.	0.75	8.74	0.30	9
s.k-NN <sup>★</sup>	L2-norm <sup>★</sup>	0.22	2.81	0.10	4
s.k-NN	LB Keogh	0.21	<b>2.29</b>	0.13	232
s.k-NN	DTW	0.23	3.05	0.15	157
K-means	L2-norm	0.26	2.65	0.08	28
K-means	DTW	0.30	3.48	0.06	31114
CONSIDERING SIMILARITY IN HISTORY CLIMATE					
2d-autocorr.	autocorr.	0.65	7.79	<b>0.36</b>	10
K-NN <sup>★</sup>	cos-sim. <sup>★</sup>	0.25	2.78	0.00	10
K-means	L2-norm	0.26	2.65	0.08	42
K-means	DTW	0.28	2.69	0.05	202

### 5.5.4 Harvesting optimization

We applied CROPi to a randomly selected timestamp – the 1000th frame from Camera RGB-2 in the 2021 split of the GSD dataset [272] – and visualized the following in Figure 5.8: observed growths until the decision moment, predicted growths using s.k-NN with L2 Norm<sup>★</sup>, changes in harvest reward (which we simply named it the “score”) over predicted growth, and suggested harvest plans under varying demands. In the third subplot, dashed lines indicate where the suggested plan advises harvesting, i.e. halting simulated growth for a non-climatic fruit like strawberry, based on a daily demand of 3 fruits from the plants that are in the view of the camera. The final result plan is illustrated in Figure 5.1.

Table 5.2: Comparison of selected methods, trained on the 2021 dataset and measured on all strawberries on a random frame in 2022.

Method	Metric	L2 ↓	DTW ↓	AC ↑
PERFECT SIGMOID FIT		0.17	1.53	0.18
PREDICTION PERFORMANCES				
Train NTP from scratch	L2-norm	0.27	2.86	0.20
Finetune GPT-2 for NTP	L2-norm	0.18	1.68	0.19
Finetune GPT-2 for NTP	CrossEntropy	0.19	2.02	0.19
s.k-NN on $G_j(t)$	L2-norm	0.20	2.35	0.16
s.k-NN on $G_j(t)$	LB Keogh	0.21	2.41	0.17
K-means on $G_j(t)$	L2-norm	0.24	2.67	0.15
k-NN on climate record	cos-sim.	0.37	3.63	0.00

As shown in Table 5.3, CROPi significantly improved the overall outcomes. Finer predictions delivered better plans and approached closer to the optimal score, which was defined using the ground-truth simulation as input. In contrast, the current harvest plan based on an empirical approach and derived from GSD annotations – yielding 2 fruits on day 1 and 8 on day 6 – performed worse overall in both harvest score and demand fulfillment compared to the other heuristics or the optimized plan.

As illustrated in the final subplot of Figure 5.1 and Table 5.3, harvests with CROPi could have provided a steadier supply with more balanced harvest schedules. In specific, when demand is high, it responds by scheduling more fruits for harvest the next day ( “+1d” ), aligning closely with observed real-world behavior; while the contract is small, it allows for slight delays to optimize fruit quality, with necessary over-harvesting to prevent rotting on the plant. By maintaining a balanced harvest distribution over time, CROPi improves overall performance and demonstrates strong potential for stabilizing supply chains through predictive, sequential decision-making.

Furthermore, the current setting of such a sparse harvesting schedule is partly due to labor constraints, as human workers cannot cover all plants efficiently. This highlights the potential of robotic harvesting, yet a simple “harvest once mature” policy remains suboptimal, particularly for maintaining a stable supply. While prediction uncertainty persists, CROPi strategically optimizes harvest timing, mitigating these challenges and demonstrating the potential of AI-driven solutions that enhance both adaptability and efficiency in precision agriculture.

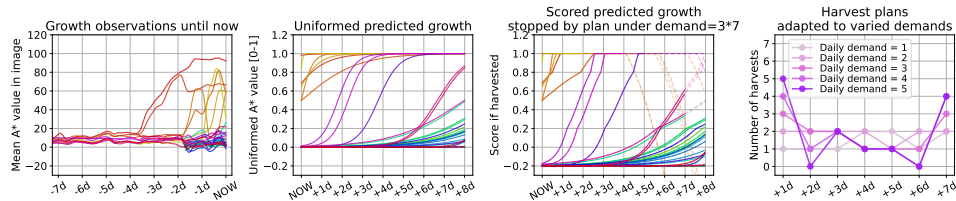


Figure 5.8: Illustrative examples using fuzzy-logic scoring with penalties, as in the third subplot of Figure 5.4. The first three subplots show observed growth, predicted growth, and scored growth, which are depicted as color-coded curves. The final subplot compares daily harvest amounts from the plans optimized over different demand levels, based on the same predictions and scoring method as in the middle subplots.



Table 5.3: Comparison of harvest plans under multiple daily demands using 2021 scenario predictions. Plans were evaluated by applying fuzzy-logic scoring with rotting penalty and over-fulfillment earnings on the simulated growths. Demand fulfillment was measured with only the basic daily demands. The results with predictions, marked with the special superscripts, were from the outputs of corresponding model settings as in Table 5.1.

Daily Demand	Harvest Method	Overall Score	Demand Fulfillment[%]
d=3	Actual harvest	5.36	28.57
	Harvest once mature	8.22	47.62
	Opt. w/ ground-truth simu.	20.63	100.00
	Opt. w/ prediction from pred.♣	18.75	100.00
	Opt. w/ prediction from pred.☆	13.41	76.19
	Opt. w/ prediction from pred.★	12.00	57.14
d=4	Actual harvest	5.36	25.00
	Harvest once mature	8.22	61.90
	Opt. w/ ground-truth simu.	21.60	78.57
	Opt. w/ prediction from pred.♣	20.01	92.86
	Opt. w/ prediction from pred.☆	14.52	64.29
d=5	Actual harvest	5.36	22.86
	Harvest once mature	8.22	<b>71.43</b>
	Opt. w/ ground-truth simu.	21.60	65.71
	Opt. w/ prediction from pred.♣	20.16	77.14
	Opt. w/ prediction from pred.☆	15.20	57.14

# 5.6 Conclusion

This chapter presents a novel end-to-end pipeline, CROPi, for optimizing fruit-level harvest decisions. In specific, CROPi uses computer vision and machine learning to quantify, track, and forecast the biological development of individual fruits, and optimization techniques to plan when and how each fruit should be harvested. Our approach integrates four key components: objective ripeness (re-)definition, growth modeling, biological development forecasting, and daily harvest planning. In a nutshell, this work demonstrates how infield monitoring can directly support data-driven harvest decisions.

In the optimization phase of our strawberry case study, we apply fuzzy-logic scoring to evaluate the plans using an objective redefinition of fruit development that is inspired by and reflects real-world practices. In this case, as illustrated in Section 5.5.1, the objective, such as generalizable ripeness definition can also function as a standalone application for consistent quality evaluations via vision systems or mobile devices. Maturity standards and market preferences, however, vary across species. In some cases, additional quality attributes such as size, firmness, sugar content, or post-harvest development must also be considered in forecasting and scoring [123, 195]. For this reason, tailored scoring criteria and corresponding data analytics are essential steps before adapting the approach to other fruits. Moreover, for risk-averse decision-makers, incorporating stochastic or robust optimization into the scoring can provide more practical and preferred alternatives.

Our demonstration includes benchmarks on three public datasets, including two for ripeness grading [76, 266] (see Section 5.4.1) and one for growth monitoring [272] (see Sec-

tions 5.4.2 and 5.4.3). Due to limited open-source data, we relied on nearby climate data [109] as a proxy for on-site microclimate information. While the weather-related model was not the top performer, its higher explainability indicates potential for improvement. In the final planning step, we used virtual demands, which could have missed details about realistic market fluctuations. More broadly, the lack of comprehensive open-source datasets is a shared concern for benchmarking how AI can be applied in agriculture [214, 272]. We hope the introduction of this novel approach also inspires the development of more comprehensive datasets to support advanced methods, such as deep reinforcement learning and predict-then-optimize strategies.

Overall, the framework demonstrates the concept of transforming diverse temporal information into a unified form suitable for optimization, by mining and harmonizing meaningful patterns from the diverse data sources. With the rapid processing enabled by the Hungarian method, CROPi supports real-time decision-making and automation, advancing smarter greenhouse management and boosting the efficiency of harvesting robots. This functionality can still be further enhanced with adaptive control methods or reinforcement learning to dynamically adjust actions in the longer term. Most importantly, as this entire end-to-end process relies solely on raw, in-field, non-destructive data, it makes the framework broadly applicable and practical for real-world use in precision agriculture and autonomous farming.

## 6

# Performance and Interaction Assessment of Model Architectures and Bivariate Smart Predict-then-Optimize

## 6

This chapter explores the impact of decision-focused learning (DFL) within a trading system applied across various fields, using renewable energy and fruit trading as examples. In this system, trades can occur through contracts or real-time markets, where the predictions of final yields are not often sufficient to lead to the most profitable contracts. We implement DFL and compare its effectiveness to traditional prediction-focused learning (PFL). Our empirical investigation, based on two real-world trading scenarios, demonstrates the novelty and generalizability of our findings, which also serve as a pre-assumption of the work presented in Chapter 5.

This chapter is based on a manuscript accepted for publication as: Junhan Wen, T. Abeel, M. de Weerd. 2025. "Performance and Interaction Assessment of Model Architectures and Bivariate Smart Predict-then-Optimize." Machine Learning Journal.



Smart “predict, then optimize” (SPO) [78] is an end-to-end learning strategy for models that predict parameters in optimization problems. Unlike minimizing mean squared error (MSE) which cares about prediction accuracies, SPO aims to ensure that predictions lead to the best possible decisions. The associated loss function, termed SPO loss, measures the decision’s regret from optimal outcomes with parameter realizations. Existing literature has demonstrated the viability of SPO, however, these studies often focus on classical optimization problems and employ a limited set of models for benchmarking.

In this study, we tackled a decision-making task inspired by real-world challenges across a wide range of neural network models. Unlike classical problems, our task requires a unique approach: collaboratively training two models to predict different variables. On top of that, one of the decision variables also affects the feasibility of the decisions, further increasing the complexity. While our implementation validates the benefits of SPO, we were surprised to find that models trained exclusively on SPO loss do not consistently attain the minimum regret. Our further investigation into hyperparameters illustrates that the well-tuned models learned very similar patterns from the feature set, irrespective of whether MSE or SPO loss was used. In other words, the change from MSE to SPO loss in training primarily affected the layer biases. Therefore, to improve the learning efficacy with SPO loss, we propose prioritizing learning feature patterns as the fundamental step. Possible strategies include using specialized neural network layers to capture deeper patterns more effectively or simply warming up by training with MSE. Specifically, a warming-up process is particularly advantageous for model(s) where the outputs are closely tied to constraints, as their prediction accuracy significantly impacts the decision feasibility. The insights are investigated empirically through two real-world trading scenarios. By leveraging datasets with diverse properties, we demonstrate the novelty and generalizability of our investigation.

## 6

## 6.1 Introduction

When machine learning models are employed to assist with downstream tasks, the impact of prediction errors can be asymmetric. In cases where the downstream task involves solving an optimization problem, this scenario is commonly referred to as a predict-then-optimize problem [78]. To enhance overall performance, it is essential to shift the focus of the learning process from prediction-oriented methods to decision-centric approaches. This paradigm shift is known as decision-focused learning (DFL) [168].

*Regret* is the core metric for evaluating decision-making performance. It describes the disparity between the optimal outcome and the outcome from the actual decision made with decision parameter estimations. To optimize the DFL process, [78] proposed training models directly on regret rather than on accuracy-based losses like mean squared error (MSE). They named this method *Smart “Predict, then Optimize”* (SPO). As such, the regret is also referred to as the *SPO loss*.

The SPO method has been widely used to tackle classical optimization problems [73, 275] and in a few abstractions of real-world cases [255, 259]. However, current research leaves two significant gaps when it comes to addressing a broader range of real-world challenges. First, most studies oversimplify decision-making by assuming only one type of unknown parameter, which can therefore be estimated with a single model or optimizer. Second,

existing research in DFL is often investigated with using only one machine learning model – typically a small neural network (NN) with fixed hyperparameters, or even simple linear models [136, 167, 246, 284]. This facilitates the explainability of the models within the DFL task; nevertheless, with the vast variety and more complex architectures of neural network models available, it is still an open question about how different architectures could enhance DFL performance.

To advance DFL research in realistic scenarios, we evaluated the performance of SPO in training diverse neural networks for a type of financial optimization task. We focused on an integrated system for both futures and real-time market trading, where products can be sold immediately or traded as futures, with transactions agreed upon in advance. Many real-world trading activities follow a pattern similar to this system, such as day-ahead bidding and last-minute trading in the renewable energy market [134, 224], the procurement of fresh agricultural products [71, 116, 222], and in financial markets where stocks, options, and futures are traded concurrently [36, 38, 57]. In these scenarios where future production cannot be perfectly controlled, a pre-set contract may result in waste from overproduction or penalties for under-delivery. Further, improper market price predictions may reduce potential profits. As a consequence, the challenge requires proper forecasts of both the *price* and the *yield*, which also acts as a constraint to the decision-making.

In short, our study investigates how different architectures and training strategies affect the models' abilities to predict and promote the downstream optimization task. Our study adopts and tailors the SPO method to train multiple prediction models that estimate decision variables (some of which act as soft constraints) respectively and collaboratively, aiming at minimizing the overall *regret* incurred by the resulting contracting decisions.

The main contributions of this work consequently are:

- We introduced a novel and realistic optimization task for product trading, of which the decision is incentivized by prices and the feasibility is bounded by yield.
- With this task, we evaluated the impact of training with *smart predict-then-optimize* (SPO), and illustrated that exclusive use of SPO loss did not improve regret consistently.
- With two real-world scenarios, we analyzed how neural networks' learning capacity fits with SPO, discovering that simpler models show more significant discrepancies of training and testing on regret, while too complex models tend to get stuck in local optima.
- The results validate that layer weights of models trained with MSE and SPO loss are similar, suggesting that training with MSE as a warm-up can enhance both effectiveness and efficiency, especially when output impacts the feasible region.

## 6.2 Related Work

### 6.2.1 Decision-focused learning

Apart from prediction-focused learning (PFL) that maximizes the accuracy of predictions, decision-focused learning (DFL) optimizes the downstream impact of decisions based on those predictions [168]. The *smart predict-then-optimize* (SPO), one of the latest DFL frameworks, introduced the decision regret as the SPO loss to train DFL models. In a linear

problem that optimizes  $\mathbf{w}$  to minimize  $\mathbf{c}^T \mathbf{w}$ , the SPO loss is written as [78]:

$$L_{\text{SPO}}(c, \hat{c}) := \max_{\mathbf{w}}(c^T \mathbf{w}) - c^T \arg\max_{\mathbf{w}}(\hat{c}^T \mathbf{w}) \quad (6.1)$$

, where  $c$  represents the decision parameter that describes a linear optimization problem in a known feasible region and  $\mathbf{w}$  is the decision variable optimized according to  $c$  (or the predicted value  $\hat{c}$ ). Training models with SPO loss have proven to be highly effective across a wide range of optimization problems [50, 169, 249, 259]. However, researchers have identified a trade-off between prediction accuracy and SPO loss during training. Thus, some suggest that accuracy may serve as a regularizer in DFL [168, 240].

Among DFL-related studies, most work has focused on a single model or a lone optimizer to demonstrate the SPO method. However, this approach falls short of realistic scenario modeling, where multiple decision variables often need to be predicted – each potentially having varied impacts on the optimal solution. For instance, the standard setting of SPO uses a known feasible region [78, 157], but the problem when the predictions may lead to decisions that violate constraints has not been well-defined yet [168]. While methods like relaxing constraints or imposing penalties improved the feasibility of solutions, they did not consistently ensure constraint satisfaction [50, 112, 136, 220, 220]. In our study, we apply SPO to a bivariate optimization problem to address the increased variable complexities that exceed those of existing setups. By training models individually, we demonstrate the effectiveness of DFL and PFL in fine-tuning models to meet specific targets.

Meanwhile, most existing works fit in learning programming formulations and simple prediction models [255]. Some applied small (two to four layers) neural networks and compared the benchmarks with classical regressors [113, 156]. Nevertheless, the choice of model architecture was always predetermined without explicitly reasoned or finetuning discussion. In fact, the family of deep learning models is much broader, with techniques ranging from traditional grid search to advanced metaheuristics for systematically exploring those hyperparameters [277]. For instance, recurrent and attention modules that are beneficial in-/extrinsic time-series analysis tasks [59, 206, 252] align well with the properties of many feature sets in DFL tasks. However, to the best of our knowledge, these neural network architectures have yet to be explored within the context of DFL.

Furthermore, many training strategies that were originally designed for PFL can be adapted for DFL tasks. For instance, [169] “warmed up” the SPO learning with a transferring learning-based idea, aiming at accelerating the training process. This approach was initially applied to demonstrate its potential for speeding up training. However, the setting of a limited number of pre-training epochs restrict the methodology’s generalizability, and other benefits remain unexplored. Despite this, the warm-up process used in PFL can easily be connected to other PFL and DFL solutions, enabling deeper discussions on how these strategies fine-tune models.

## 6.2.2 Optimizations constrained by contracting decisions

At present, there exist many situations where the same product is traded in multiple markets, with varying time lags between the trading and transaction dates. For example, in the renewable energy sector, prices fluctuate due to shifts in supply and demand [32, 134].

To manage this volatility, traders commonly use supply contracts to stabilize and secure transactions and turn to real-time markets for last-minute adjustments [202, 289]. A comparable strategy is found in the trade of perishable agricultural products, commonly referred to as *contract farming* in many countries [71, 116, 222]. Price stability encourages growers to sell most of the harvests through contracts, with surplus typically directed to local markets. However, imbalances in negotiation power – particularly between small suppliers and large buyers such as supermarkets – can lead to inefficiencies, which often manifest as either over-supply or under-supply [81, 189, 234].

Since contracts require advance commitments while the real-time market has higher uncertainty but lower capacity, producers face the challenge of determining how much future production to allocate to contracts. Existing studies estimate the profitability under such situations by statistical models such as mean-variance and probability density distribution models [53, 137, 224]. To optimize contracting decisions, common methods include stochastic optimization and worst-case analysis models [155, 202, 216, 224]. However, these models generally prioritize accuracy in predicting future production and real-time prices, often overlooking the specific consequences of under- or over-supply, such as contract-breaching fines [32]. Consequently, such predictions should (also) be evaluated based on their ultimate impact, which underscores the importance of the DFL concept.

## 6.3 Research Questions

## 6

Previous works have demonstrated the advancement of using SPO loss in training models for decision-focused learning (DFL) purposes. However, there is no common understanding of what machine learning models we need to use and which model is suited for which problems. To enhance the integration of the SPO method with a broader range of deep learning tools, we explore a variety of neural network architectures with diverse learning capabilities and assess the PFL and DFL performance. In essence, this research is designed to answer the following questions:

- RQ1.** How does the learning capacity of models influence PFL and DFL?
- RQ2.** How do MSE and SPO loss functions influence the fine-tuning of models?
- RQ3.** How to optimize the DFL performance under a given model architecture?
- RQ4.** How can training strategies be tailored to align with the roles of the target (decision) parameters being predicted?

## 6.4 Methodology

In this study, we comprehensively investigated the influence of model design and training setup in the implementation of the end-to-end training strategy *smart predict-then-optimize* (SPO) [78]. We addressed our research inquiries by benchmarking against realistic variants of a portfolio decision-making scenario. The objective of the use case is to decide how much future production should be committed to a given trading contract while both the (other) market's price and the future yield are unknown. Under this scenario, the yield directly influences the scale of the decision and the price mainly incentivizes the decision towards one of the trading channels. This section outlines our methodology from

four key perspectives: model selections, metric choices, loss function decisions, and how we came up with the improvement strategy.

**Models:** To explore DFL with a border range of models, we exploited neural networks with different depths of multi-layer perception (MLP) layers, with a recurrent neural network (RNN) layer, with pair-wise self-attention module (ATT) after the input layer, plus one with just one linear layer to represent a linear regressor. RNN and ATT were chosen because the features are extrinsic time-series data. In the use case specifically, two models were trained to predict the market price and production. They were trained collaboratively from scratch with specified optimizers. By employing diverse training settings and strategies, we empirically studied the effectiveness of PFL and DFL on models with different learning capacities.

**Metrics:** The primary metric for evaluating model performance is the average *regret* on the testing set, which reflects the expected “unearned” monetary units from real-world decisions. To assess real-world impact, we compared the *profit* distributions under variable predictions with optimal decision outcomes under parameter realizations. We also analyzed the mean squared error (*MSE*) of predictions, as it indicates the generalizability of the model and the training strategy. We used the mean error (*ME*) of predictions to demonstrate the bias of models. Additionally, we discussed the performance related to *timing*. To compare how PFL and DFL fine-tune the models, we measured the *cosine similarity* of corresponding layers in the trained models, which reveals how each method influences the weight and bias of the layers.

**Training losses:** Our study began with models trained exclusively on MSE or SPO loss. We then refined the experiments by training the models using a weighted average of both MSE and SPO loss, i.e. the loss  $L = w_1 \cdot \text{regret} + w_2 \cdot \text{MSE}$ , s.t.  $w_1 + w_2 = 1$ . By varying the ratio  $w_1 : w_2$  from 0 to infinite, we evaluated how the models are progressively tailored in terms of PFL and DFL goals. We applied two independent ratios for the price and yield prediction models respectively, so as to explore specific preferences for each model’s training and optimize the learning efficacy.

**Improvements:** By comparing the final weights of models and investigating the metrics during the training process of PFL and DFL, we decided to “warm up” the DFL training from models converged on PFL process. We compared the performance with the other models that were trained by DFL from scratch, and with models trained under diverse numbers of PFL epochs, so as to generalize the methodology. On top of that, we conducted such initialization on only one of the models to further explore whether there is a core model to start with or if the collaboration essentially affects the training. The in-depth exploration suggests the optimal design of learning settings, referring to the outputs’ specific functionalities in the downstream optimization task.

## 6.5 Problem and Task Formulation

### 6.5.1 Background

When the future production of a product cannot be perfectly controlled, such as with agricultural products or renewable energy, maintaining steady relationships with downstream

partners can be challenging. To manage the risks associated with forecast errors, traders often diversify their strategies across markets that differ in how far in advance decisions must be made [260]. For example, some contracts require early commitments, while real-time markets are typically smaller and more volatile [116, 134]. Poor forecasts can lead to overproduction and waste, or shortfalls that result in monetary penalties and damage relationships with downstream buyers. In addition, inaccurate predictions about market conditions might also cause producers to miss more profitable sales opportunities.

In related studies, the profitability is estimated by statistical models, e.g. by a mean-variance model or probability density distributions [53, 137, 224]. However, these methods can lead to high penalties when the decisions are diverged from the predictions. To avoid such, growers, for example, may opt for conservative yield estimations [189, 234]. Thus, it is more insightful to evaluate predictions based on their impact on decisions rather than on accuracy alone. This approach aligns well with the DFL strategy.

We chose this type of use cases because it represents a common scenario across many financial systems, featuring a straightforward and distinct goal that serves as a basis for experimenting with various models. Even so, the problem can be more complex and pose greater challenges than many existing DFL works, as it involves predicting multiple parameters (prices, production, etc.) and the available feature set may not always have a linear relationship with the target outputs.

## 6.5.2 Problem characteristics

In this chapter, we investigated a use case where the goal is to maximize the overall profit for each batch of production by determining the best trading strategy within an integrated contract and market system. For the decision, we specify how much of a future production batch will be sold through contractual agreements at a predetermined price, and how much will be reserved for sale in the market, where prices fluctuate in real time. Generally, the closer the time to the transaction, the higher accuracy of prediction can be achieved. Nevertheless, real-time demand and price have significantly higher variations than trading through futures or options [58, 222, 257] – sometimes there can be barely any demand and thus significantly low prices, but at other times their profitability can surge in response to sudden demand or an unforeseen supply shortage.

In specific, this problem scenario has the following characteristics:

1. We act as a price-taker in all trading activities.
2. At the time of decision-making, the contract prices  $P_c$ <sup>1</sup> and transaction time  $t$  are known, and both the final production volume (shortened as  $Y$  for yield) and the (real-time) market price  $P_r$  can be predicted based on current trends.
3. Once production is completed, we first fulfill our contract obligations by delivering the agreed-upon amount, or as much as we can. The remaining product is then sold in the market at the prevailing price.

<sup>1</sup>We use bold fonts when introducing new parameters. After that, we use roman font for decision parameters such as  $P_c, P_r, Y$ , and calligraphic font for the decision variable, i.e. the contracting amount  $\mathcal{D}_c$  and  $\tilde{\mathcal{D}}_c$ .

4. The volume of contract-based trading is constrained by the capacity of the contract and the actual yield of our production. Moreover, contracted buyers generally do not accept purchases beyond the committed amount at decision-making time.

### 6.5.3 Objective function and specific SPO framework

In this use case, the optimization objective is the overall profit of a batch of production, denoted as  $\text{Pft}(\mathcal{D}_c, T_c, T_r, Y)$ , wherein  $\mathcal{D}_c$  is the contracting decision,  $T_c$  means the contract transaction,  $T_r$  means the market transaction, and  $Y$  represents the final production.

By definition, the profit is the difference between the total revenue, i.e.  $P_c \cdot T_c + P_r \cdot T_r$ , and the costs. In line with established practices in the field [81, 145], we considered linear competitive market models to describe market capacities, i.e. the maximum volumes for contract trading is  $T_{c,\max} := a_{ct} \cdot P_c + b_{ct}$ , and in the real-time market is  $T_{r,\max} := a_{rt} \cdot P_r + b_{rt}$ . Consequently, to focus solely on the parameters derived from the datasets, the objective function can be formulated as  $\text{Pft}(\mathcal{D}_c, P_c, P_r, Y)$ .

According to the trading mechanism, at the transaction date, the actual transaction amount with the contract is the (original) decision bounded by the actual production and the contract capacity:  $T_c = \min\{\mathcal{D}_c, Y, T_{c,\max}\}$ . The market transaction is bounded by the remaining production and the market capacity:  $T_r = \min\{Y - T_c, T_{r,\max}\}$ . Given the real-time price and production are unknown at the time of contracting, the decision is actually made based on their predicted values  $\hat{P}_r$  and  $\hat{Y}$ . Hence, the best decision that we can make is  $\hat{\mathcal{D}}_c = \arg\max_{\mathcal{D}_c} \text{Pft}(\mathcal{D}_c, P_c, \hat{P}_r, \hat{Y})$ . Consequently, later at the transaction execution date, the actual contract transaction is  $T_c = \min\{\hat{\mathcal{D}}_c, Y, T_{c,\max}\}$ . In this way, the yield  $Y$  serves as a soft constraint of the optimization problem.

The costs consist of (i) the internal production cost that depends on the production:  $a_{c1} \cdot Y$  and (ii-iii) the external cost from improper decisions: (ii) if the actual production falls short of the contracted amount, i.e.  $\mathcal{D}_c > Y$ , we pay a penalty for every under-delivery unit  $a_{c2} \cdot \text{pos}[\mathcal{D}_c - Y]$  and we trade to the contract with  $T_c = Y$  instead of  $T_c = \hat{\mathcal{D}}_c$  because  $T_c = \min\{\hat{\mathcal{D}}_c, Y, T_{c,\max}\}$ ; (iii) or if the overall production exceeds the contracted amount and the market's capacity, the waste is accounted for in financial terms:  $a_{c3} \cdot \text{pos}[Y - T_c - T_r]$ . Note that in this specific context, the penalty for infeasible decisions is smoothly incorporated by the under-delivery costs.

Hence, the objective function, i.e. the profit  $\text{Pft}$ , is written as:

$$\begin{aligned}
 & \text{Pft}(\mathcal{D}_c, P_c, P_r, Y) \\
 &= P_c \cdot T_c + P_r \cdot T_r - a_{c1} \cdot Y - a_{c2} \cdot \text{pos}[\mathcal{D}_c - Y] - a_{c3} \cdot \text{pos}[Y - T_c - T_r] \\
 &= P_c \cdot \min\{\mathcal{D}_c, Y, (a_{ct} \cdot P_c + b_{ct})\} + P_r \cdot \min\{Y - \min\{\mathcal{D}_c, Y, (a_{ct} \cdot P_c + b_{ct})\}, (a_{rt} \cdot P_r + b_{rt})\} \\
 &\quad - a_{c1} \cdot Y \\
 &\quad - a_{c2} \cdot \text{pos}[\mathcal{D}_c - Y] \\
 &\quad - a_{c3} \cdot \text{pos}[Y - \min\{\mathcal{D}_c, Y, (a_{ct} \cdot P_c + b_{ct})\} - \min\{Y - \min\{\mathcal{D}_c, Y, (a_{ct} \cdot P_c + b_{ct})\}, (a_{rt} \cdot P_r + b_{rt})\}]
 \end{aligned} \tag{6.2}$$

$$\tag{6.3}$$



By capturing the core concept of the original SPO loss as in Equation (6.1) and given that the contract price  $P_c$  is known at the moment of decision-making, we derived the specific SPO loss for the market price and yield prediction  $\hat{P}_r$  and  $\hat{Y}$  as:

$$\begin{aligned} L_{SPO}(P_r, \hat{P}_r, Y, \hat{Y}) &:= \max_{\mathcal{D}_c} \text{Pft}(\mathcal{D}_c, P_c, P_r, Y) - \text{Pft}(\tilde{\mathcal{D}}_c, P_c, P_r, Y) \\ &= \max_{\mathcal{D}_c} \text{Pft}(\mathcal{D}_c, P_c, P_r, Y) - \text{Pft}(\arg\max_{\mathcal{D}_c} \text{Pft}(\mathcal{D}_c, P_c, \hat{P}_r, \hat{Y}), P_c, P_r, Y) \quad (6.4) \end{aligned}$$

### 6.5.4 Data

To illustrate the decision problem at hand and to assess the effectiveness of the proposed methodology in realistic settings, the experiments of the chapter are conducted as case studies. The experiment section mainly focuses on the decision-making performance of a wind farm that participates in the day-ahead and real-time market, using the energy price and consumption data sourced from [121] (refer to as the “wind” case).

The decisions of trading in the day-ahead market, which we identify as the channel for “contracting”, must be finalized before the bid clearing deadline at 12 pm [5, 58, 289]. When there is a shortage of production, we may opt to purchase additional electricity to fulfill our committed amount; nevertheless, in instances of unplanned overproduction, the cost of finding immediate consumers can be prohibitively high [58]. Consequently, the expense incurred from overproduction significantly outweighs that of underproduction, underscoring the importance of employing DFL in predicting the unknowns. The dataset covers 34,613 timestamps, collected hourly from 2015 to 2018. Due to the large computational load of DFL, we used the data from every tenth hour for exploitative studies. The feature set had 143 attributes in total, consisting of local weather records from the past week, trading information from the past four realizations, and the contract price for the current decision round. The price was scaled by 100 from the original monetary unit (euro), and the generation was scaled by 10,000 to facilitate neural network training. Detailed settings, exploration, and accessibility of the data can be referred to the appendix.

In fact, in 90% of the instances in our primary dataset, the real-time market price exceeds the daily market price, which serves as our contract price. To address this bias, we performed an ablation study that ensures an equal likelihood of either price being higher than the other, and we refer to this as the “wind-50” case.

In addition, we curated a dataset for another use case involving a similar trading optimization problem, but with different characteristics from the wind trading scenario. In this case, we consider the role of a tomato grower, where the task is to decide how much yield to commit through farming contracts [71, 189]. The available price data, sourced from [33], is reported daily, though not consistently for every single day. Given that prices are provided in upper and lower limits but without specifying the trading channel, we considered two scenarios: in the first, the upper limit is seen as the contracting price and the lower limit as the market price (referred to as “tomato-c” as the contract price is higher); in the second scenario, we reversed this arrangement, ensuring that the contract is not more favored than the market (referred to as “tomato-m”).



We used similar attributes to form the features for the two tomato cases. Nonetheless, since the decision-making time and available data are much sparser, we used the daily temperature record in the past 28 days of the decision date and the realized cases of the past four available records within a month. Together with the current contracting price, the feature set had 41 attributes. In total, we had 369 data points, collected from January 2014 to August 2023. For both cases, we scaled the features by constants to facilitate the training. These two cases, together with the “*wind-50*” case, compared the impact of DFL on yield ( $Y$ ) and market price ( $P_r$ ) predictions under more diverse scenarios. The decision performances on these three cases are discussed in Section 6.6.5.

## 6.6 Experimental Results

In this section, we present experimental results on investigating and refining the decision-focused learning (DFL) process with SPO loss by primarily the *wind* case, so as to gain deeper insights and address the research questions outlined in Section 6.3.

Our exploration begins with the examination of multi-layer perception (MLP) models with varied numbers of layers. This includes a one-layer model without activation, which essentially functions as a linear regressor. Additionally, we analyzed the performance implications of incorporating a fully-connected recurrent (RNN) layer or a pair-wise self-attention (ATT) layer alongside the MLP layers. All models started with 256 neurons and ended with 512, except for the linear regressor, which had only one layer of 256 neurons. Dropout layers were added after every two MLP layers. For model trained by DFL, the outputs of the neural networks flowed into a CVXPY layer [69] to compute and back-propagate the SPO loss. We trained both the price- and yield-prediction models using identical model architectures and input features. Unless specified otherwise, the models were trained together from scratch. Hyperparameter fine-tuning is further discussed in the appendix.

In this section, we present results from models trained by stochastic gradient descent (SGD) optimizers with a  $2e-3$  learning rate, a  $1e-5$  weight decay regularization, and a maximum of 1,000 epochs with early stopping. The dataset is divided into 7:1:2 for training, validation, and testing. All experiments were run with five fixed random seeds unless otherwise noted. The experiments were carried out on an NVIDIA A40 GPU.

### 6.6.1 Model architecture impacts performance and training-testing discrepancy on regret

This subsection addresses the first research question regarding how the learning capabilities of MLP models affect outcomes. Figure 6.1 depicted the final regrets, resulting from decisions optimized using parameters predicted by models of different depths. The plot on the left compares the total regret across three cases: when neither model, only one, or both models were trained using the SPO loss. The four smaller plots on the right provide a closer look at regret under various combinations of training losses. Regret results from models trained with both MSE and SPO loss are included to link the findings between PFL and DFL approaches.

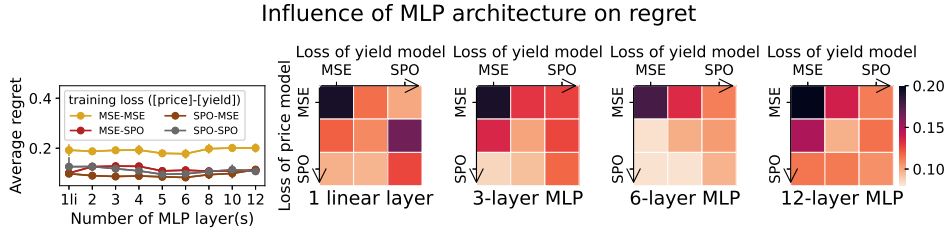


Figure 6.1: Regret of models varies with MLP layer numbers and loss function combinations. The left plot compares regret over depths for models trained on different loss functions. The right four plots compare regrets under different loss functions and model architectures. For the right plots, the x-axis label means the loss for yield prediction training and the y-axis label is that for price. The arrows on the x- and y-axis indicate the increasing involvement of SPO in the loss function. All plots share the same color scale, shown on the right. The results in the middle boxes used loss =  $0.5 \cdot \text{MSE} + 0.5 \cdot \text{SPO}$ . Results are collected over five seeds.

As illustrated in the first subplot, models trained both by PFL exhibited the poorest decision-making performance, underscoring the necessity of integrating SPO loss into the training process. The subsequent four colored plots further reveal that decision performance generally improves when the price model training includes more SPO loss. However, this trend isn't consistently seen when increasing SPO loss in the yield prediction model's training. In other words, the lowest regret doesn't always occur when all models are trained exclusively with SPO loss. This inconsistency is especially clear in shallower models, while deeper models tend to perform worse overall, likely because they get stuck in local optima.

6

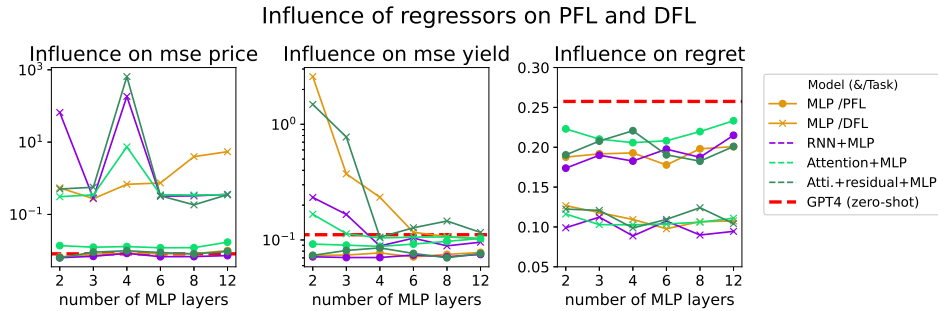


Figure 6.2: MSE of the price (left), yield (middle) prediction models, and their resulting regret (right). The x-axis represents the number of MLP layers. Lines are color-coded by neural network architecture, as indicated in the legend. Dot markers show performance for models trained under PFL, while 'x' markers indicate DFL-trained models. Except for the zero-shot performance of GPT-4, results are averaged over five seeds.

Next, we compare the performance among four types of neural networks: MLP, RNN-MLP, ATT-MLP, ATT-MLP and with MLP residuals. Figure 6.2 presents the average MSE and regret from the models with varying numbers of MLP layers and different loss functions. The prediction accuracies of the models are close to state-of-the-art – for example, comparable to the method by [188] – and can improve slightly with more complex models. This suggests that all architectures are suitable for the learning tasks in this scenario. However, these gains in accuracy only modestly reduce final regret to a limited extent. In

contrast, training models with DFL produces more substantial gains in decision-making performance. As shown in the third plot of Figure 6.2, this advantage is consistent across all architectures, highlighting that optimizing directly for regret is more effective than simply increasing model complexity while using the same loss function.

We also analyzed the correlations of Kullback-Leibler (KL) divergence between regret and the MSEs of the price and yield predictions across five test sets. Our results show that regret is moderately positively correlated with errors in yield prediction (0.61), suggesting that yield prediction quality does influence the regret. In contrast, regret has almost no correlation with price prediction errors (-0.01), indicating that perfect accuracy in price prediction is not always necessary for DFL purposes. Interestingly, the PFL performances of the two prediction models show a slight negative correlation (-0.37), which further supports the idea that these models play different roles in the optimization task.

## 6.6.2 Models learn alike feature patterns from PFL and DFL

To better understand the mechanisms behind PFL and DFL training, we analyzed the cosine similarity between layer parameters of models trained with different loss functions. Figure 6.3 shows the results for models with six MLP layers, which was the best-performing architecture among models in Figures 6.1 and 6.2. We observe notable similarities in the weights of MLP layers, while the differences appear mainly in the layer biases<sup>2</sup>. These results suggest that the divergence between PFL and DFL primarily affects the bias terms, with limited impact on feature representations. This observation motivates our strategy of pre-training on PFL before fine-tuning with DFL, as shown in the final plot of Figure 6.4. The impact is further discussed in Section 6.6.3.

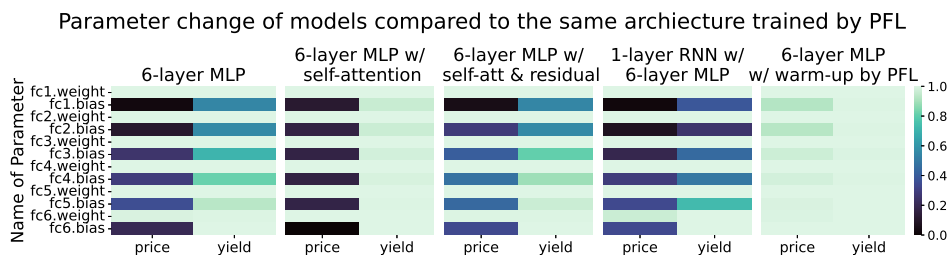


Figure 6.3: Cosine similarity of layer parameters between models with six MLP layers (“fc”) trained with SPO and MSE. The final plot compares a PFL model with a DFL model that was initialized with the PFL model’s weights (referred to as a “warm-up”), while the other plots compare PFL-only models to DFL models trained from scratch. Layer indices are shown on the y-axis. All plots share the color bar on the right, ranging from 0 to 1.

Interestingly, as shown in Figure 6.4, with increasing involvement of SPO loss during training, the regret performance of ATT models changes similarly to a much deeper MLP model, such as that with 12 layers as depicted in Figure 6.1. Both architectures show weaker DFL performance compared to others, suggesting that models with very high learning capacity may be more prone to getting stuck in local optima. As illustrated in Figure 6.3, the price prediction component in these models appears to be overly influenced by DFL, which may

<sup>2</sup>Similar trends were found in the self-attention and RNN layers, detailed in the appendix.

contribute to the performance decline. One potential solution is to use MLP outputs as residuals alongside the self-attention layers. As demonstrated in Figure 6.3 and Figure 6.4, this adjustment helps produce more balanced prediction biases and improves performance in both tasks. Furthermore, as indicated in Figure 6.2, these models also achieve lower MSEs, indicating improved generalizability.

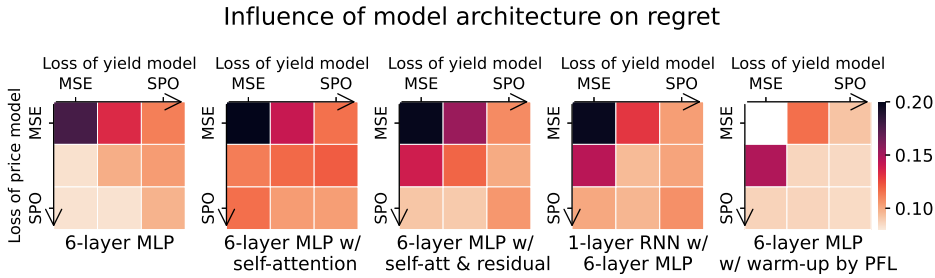


Figure 6.4: Regret from decision-making, using parameters from models with varied architectures and trained on different loss functions. The models are aligned with those used in the layer parameter comparisons (figure 6.3). The x- and y-axis follow the same configuration as in figure 6.1. All plots share the same color scale as on the right. The upper-left corner of the final plot is left empty, as it represents again a purely PFL-trained model.

### 6.6.3 Pre-training mitigates discrepancy and reduces regret

Motivated by the strikingly similar feature patterns learned by PFL and DFL, we investigated using PFL as a form of pre-training – or “warm-up” – for the DFL process. As illustrated in Figure 6.4, incorporating SPO loss after such a warm-up significantly reduced the performance discrepancy between training and testing on regret, leading to improved overall performance. Notably, when comparing models trained with identical loss functions from scratch, the improvement is most pronounced when the model is trained exclusively with SPO loss. Further analysis in Figure 6.3 reveals that this warm-up strategy results in stronger alignment of layer-wise biases with the original PFL model, compared to models where DFL is trained entirely from scratch. This suggests that warm-starting DFL encourages smoother and more consistent bias adaptation, which likely improves the models’ generalizability.

It is important to note that in our problem, *yield* naturally serves as a soft constraint of the objective function, whereas *price* primarily incentivizes decision direction, rather than its magnitude. This difference affects how DFL introduces bias into the two prediction models. Since decision feasibility is highly sensitive to changes in the yield model, preserving its pre-trained structure is critical. With warm-started DFL, the yield model changes very little, helping to avoid harmful bias and supporting stable decision-making. Conversely, the price model can tolerate – or can even benefit from – certain bias introduced during DFL, as it helps align predictions with task-specific goals. Still, compared to training from scratch, the bias introduced by performing DFL as a fine-tuning process remains more moderate, reflecting a better balance between task adaptation and generalization.

Considering that price and yield function differently, we conducted comparative experiments to assess the effectiveness of selectively warming up the training of each model. As

is shown in Figure 6.5, initializing the yield predictions model with pre-trained weights proves essential for solving the optimization problem effectively, as it ensures that decisions are made under more realistic constraints, which contributes to the lower regret observed. These findings reinforce that, for the yield model, preserving accuracy is more important than inducing bias aligned with the task. In other words, an accurate estimation of feasible regions provides a more reliable starting point for the following DFL phase.

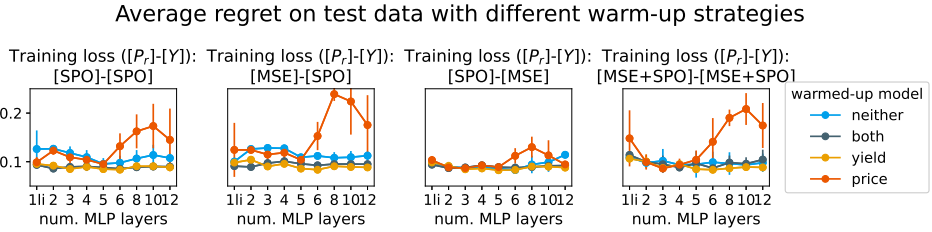


Figure 6.5: Comparison of the effect of the pre-training strategy by the regret, of models with different training losses. The color codes show which model(s) was/were pre-training on the MSE. The x-axis shows the model depth, with “1li” referring to linear regression.

In the above experiments, the warm-up phase lasts until the model converges over 20 epochs based on MSE. As illustrated in Figure 6.6, the impact of warm-up on final performance also depends on how well the model is pre-trained. To further explore this, we compared our convergence-based, dynamic approach against using a fixed warm-up of 6 epochs, as used by [169]. While a short fixed warm-up can still offer some benefit, our results suggest that a convergence-based schedule with sufficient patience contributes to more robust and consistently better performance.

Meanwhile, implementing a warm-up process for both models reduces the average training time from 1.23 hours to 0.79 hours<sup>3</sup>. This time-saving effect is particularly notable when only the yield prediction model undergoes the warm-up process. In contrast, warming up only the price prediction model does not result in a significant reduction in computational time.

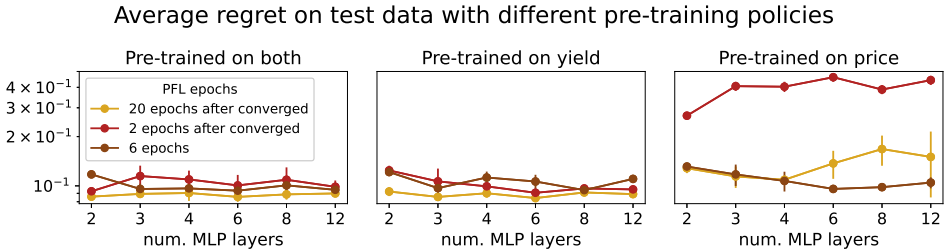


Figure 6.6: Resulting regret of models warmed up differently. The titles indicate which model(s) was/were pre-trained. The color shows the periods of pre-training.

<sup>3</sup>Averaged across all models in Figure 6.5 and all seeds. More comparisons can be referred to in the appendix.

### 6.6.4 The sub-problems unveiled as classical DFL tasks

By using the ground truth of one parameter and implementing DFL with SPO loss on the other, we simplify the real-world problem to a more ideal and classical DFL task. From the middle plot in Figure 6.7, it is clear that SPO is a powerful solution when the yield is known. In this case, where all constraints are fixed, the problem setting closely resembles the one proposed by Elmachtoub and Grigas [78], underscoring the strength of SPO under such conditions. Thus, applying DFL to the task of price prediction represents a meaningful and novel advancement over prior approaches.

However, the scenario changes when prices across all trading channels are known and the prediction target shifts to yield. In this setting, yield directly influences key operational outcomes such as overproduction costs and supply shortage penalties. As a result, achieving high prediction accuracy becomes even more crucial. This shift is illustrated by the third subplot in Figure 6.7: the varying gaps between the three solid lines further emphasize how the benefits of DFL depend significantly on the decision parameters involved and their specific roles within the optimization.

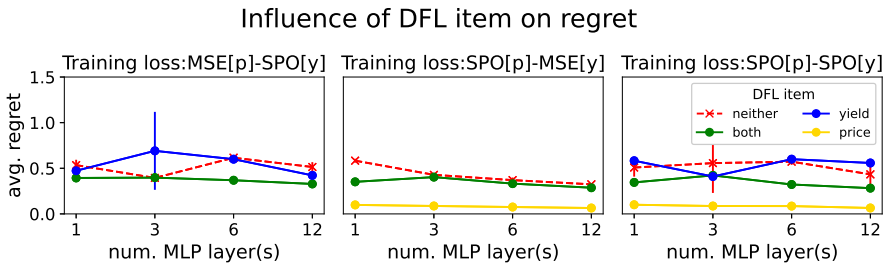


Figure 6.7: Regret of MLP models when one parameter is considered as known in the DFL task, compared to previous PFL and (both-)DFL models' performances. The subtitles indicate the specific loss function of the yield ([y]) and price ([p]) prediction model. The x-axis shows the depth of the MLP model.

Further, when reviewing the training progress in detail, we observed that gradient vanished several times when training the price prediction model solely on the SPO loss. This issue, which is known as the zero-gradient problem in several DFL studies [168], occurred because *price*, in this context, functions purely as a decision-driven parameter. Consequently, without incorporating a regularization strategy or a warm-up phase using MSE during training, the model can become significantly biased, as demonstrated in Figure 6.3.

### 6.6.5 Real-world impact

In previous sections, we focused on model performance across training losses and strategies, using the “*wind-ori*” dataset as an illustrative example due to its realistic and sufficient data. Here, we evaluate the regret from 3-layer MLP models across the four scenarios introduced in Section 6.5.4: the original wind farm case (“*wind-ori*”), a modified version with more balanced price signals (“*wind-50*”), and two tomato-grower scenarios (“*tomato-c*” and “*tomato-m*”), where underproduction incurs heavier penalties than overproduction, contrasting with the wind farm cases. We also consider the real-world impact of DFL by

evaluating mean error (ME) across the four cases and the profitability in the “*wind-ori*” scenario. A summary of the resulting regrets under different learning settings are depicted in Figure 6.8. To better understand the real-world implications of applying DFL, we also evaluate the mean error (ME) across these four cases, as described in Table 6.1.

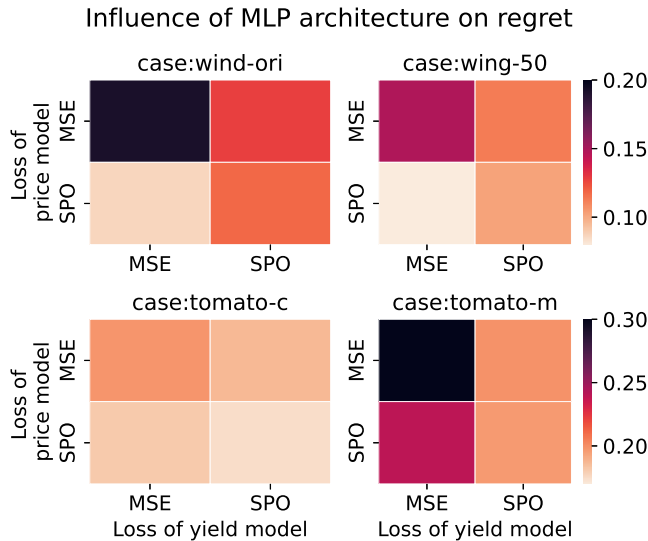


Figure 6.8: Average regret of price and yield prediction models in different experiment cases. The loss function is indicated in the x and y labels respectively. The wind cases share the color bar in the first row and the tomato cases share that in the second row on the right.

Table 6.1 shows that DFL intentionally introduces bias in yield and price models to account for asymmetric penalties: yield predictions are skewed to avoid costly underproduction, while price predictions exaggerate contract-market price differences to favor more profitable decisions. These biases reduce overall regret compared to models trained with MSE alone, which suffer from symmetric error distributions that hurt downstream tasks.

Nonetheless, as Figure 6.8 illustrates, regret is a challenging loss for models to optimize directly. This difficulty is especially evident when training jointly on regret for both tasks, resulting in a noticeable gap between training and testing performance. Interestingly, we find that lower bias in yield prediction often aligns with reduced regret in these cases. This suggests that carefully controlled bias in DFL can improve prediction models and lead to better downstream outcomes.

Figure 6.9 shows the profiting performance of selected models and two more case-specific baselines: one used the optimal decision from the latest parameter realization (“last-round decision”), and another trained models using MSE weighted by under- and over-supply costs, i.e.  $a_{c2}$  and  $a_{c3}$  in Equation (6.3) (“MLP: weighted-MSE”). These baselines, along with the two blue boxplots, represent traditional two-stage approaches: first the forecasting models are trained independently, and then decisions are optimized based on their predictions. Moreover, while weighting MSE helps reduce extreme losses, it falls short

Table 6.1: ME of price and yield predictions in different experiment cases. The loss function is indicated by the 2<sup>nd</sup> and 4<sup>th</sup> columns.

Case	Price		Yield	
	Training Loss	Testing ME	Training Loss	Testing ME
tomato-c	MSE	0.26	MSE	-0.07
	MSE	0.23	SPO	<b>0.34</b>
	SPO	<b>-1.31</b>	MSE	-0.07
	SPO	<b>-1.32</b>	SPO	<b>0.26</b>
tomato-m	MSE	1.10	MSE	-0.02
	MSE	1.01	SPO	0.02
	SPO	<b>-1.18</b>	MSE	-0.04
	SPO	<b>-1.32</b>	SPO	-0.04
wind-ori	MSE	-0.08	MSE	-0.02
	MSE	-0.01	SPO	<b>0.50</b>
	SPO	<b>1.74</b>	MSE	-0.05
	SPO	<b>0.76</b>	SPO	0.05
wind-50	MSE	-0.08	MSE	-0.03
	MSE	-0.22	SPO	<b>0.47</b>
	SPO	<b>2.11</b>	MSE	-0.05
	SPO	<b>-0.23</b>	SPO	<b>0.40</b>

of matching the profit levels achieved by DFL-trained models. This suggests that simply improving the forecasting model has limited impact unless the prediction is aligned more directly with the downstream objective.

Across all boxplots, the models with DFL applied on top of PFL-pretrained models (cyan boxes) stand out by both raising the upper bound of achievable profits and reducing losses in the lower tail. This indicates that their improved performance comes not just from higher average returns, but from a more favorable overall distribution – achieving stronger gains while limiting downside risk. The stability introduced by PFL helps maintain reliable yield predictions, which in turn supports the SPO algorithm in enforcing constraints more effectively. This enables DFL to better guide the prediction of decision parameters and ultimately leads to more robust decision-making outcomes.

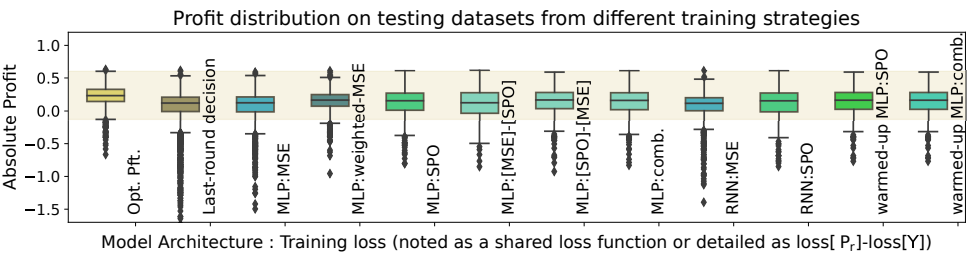


Figure 6.9: Profit distribution in the “wind-ori” case, with a 6-layer MLP model. The [architecture:training loss] configuration is indicated next to each box plot. The distribution plots are color-coded to distinguish the loss functions and training strategies. Profits from optimal decisions based on parameter realizations are shown in the yellow box, with the lighter yellow band serving as a value reference for other distributions.



## 6.7 Discussion

In this study, we investigated the comparative effectiveness of prediction-focused learning (PFL) and decision-focused learning (DFL) within the “*smart predict-then-optimize*” (SPO) framework. We empirically assess model performance through a real-world task: making selling decisions in an integrated contract and (real-time) market trading system. Our research extends existing DFL studies by (i) incorporating diverse neural network architectures and exploring the impact of their learning capabilities, and (ii) addressing a more intricate decision-making problem, where models collaboratively predict two types of variables: market *price* as a decision parameter and *yield* as both a decision parameter and a soft constraint.

Our benchmarks underlined a notable performance discrepancy: Models trained solely on the *regret*, which is also known as the SPO loss, do not consistently lead decisions on the test set to achieve the lowest regret. This inconsistency appears across various model architectures and is especially pronounced in simpler models. Although we have explored a variety of network architectures, our study covers only a small slice of the landscape of deep learning. Future work aiming to discover the optimal network depth and architecture in more problem settings may benefit from metaheuristic methods such as those proposed in [139] and [8].

### 6

In terms of the efficacy of DFL and PFL, our investigation into model hyperparameters demonstrates that although models trained with Mean Squared Error (MSE) and SPO loss exhibit different biases, they tend to converge on similar layer weights and self-attention patterns. Considering that PFL is more computationally efficient, it gives a clear motivation for enhancing DFL by starting from models pre-trained with MSE. This warm-up approach offers three main benefits: (a) narrowing the training-testing performance gap by boosting DFL effectiveness; (b) reducing computational cost; and (c) improving generalization by regularizing the bias introduced during DFL training.

In our specific use cases involving decision making as a wind farm operator or a tomato grower, our findings highlight two key insights: DFL can change price predictions compared to PFL, while accurate yield prediction is especially important when the yield serves as a constraint in the decision-making process. This difference comes from the distinct roles that price and yield play in the objective functions. In this way, our findings offer clearer insights into the SPO method’s applicability and highlights the importance of modeling constraint-related variables with care. Similar advantages were also evidenced in other constrained decision-making scenarios, such as safe reinforcement learning [288].

In all case studies, the same input features were used for both prediction tasks, naturally motivating the use of multitask learning (MTL) to improve learning and data efficiency. Prior work [241] has shown that MTL with DFL can be especially effective for smaller datasets. While we do not evaluate MTL in the wind farm setting due to its larger dataset size, it presents a promising direction for sparser cases like the tomato price dataset discussed in Section 6.6.5.

# 7

## Discussion

In this thesis, we present four studies that systematically develop data-driven solutions for a challenge investigated across three interconnected levels: the fruit, the greenhouse, and the market. The work begins at the fruit level (Chapters 3 and 4), focusing on predicting individual fruit characteristics. This analysis is then scaled up to the greenhouse level in Chapter 5 to support harvest decision-making. Finally, the investigation is elevated to the market level in Chapter 6 to analyze the overarching strategic and economic context. This multi-level approach is supported by a combination of unique datasets curated for this research (presented in Chapter 2) and publicly available market and climate data. Ultimately, this research provides a holistic and coherent solution to the initial real-world problem: optimizing the supply and harvest planning of a soft-fruit supply chain.

This discussion chapter provides a comprehensive overview of the thesis from two distinct perspectives. First, we present detailed, question-by-question responses to the eight research questions that form the foundation of the study. These responses are drawn directly from the findings presented in the published papers, which constitute the main, interconnected, chapters of the thesis. Second, we synthesize three overarching insights that go beyond the scope of the individual research questions. These insights integrate and contextualize the key findings, highlighting their significance to the broader field of machine learning (ML) and its applications on decision-making tasks in the soft-fruit supply chain. Accompanying these insights, we identify potential directions for future research, underscoring areas where further exploration can extend the contributions of this work.

### 7.1 Answers to Research Questions

The four paper-based chapters, Chapters 3 to 6, answered the eight research questions framed in detail in Section 1.2. Our contributions span the entire pipeline, from data acquisition and modeling to operational optimization and strategic decision-making. We began by developing methods to monitor and assess the fruits when they are still in field, specifically, by investigating Multiple-Object Tracking (MOT) algorithms to keep track of the fruits and designing ML models to predict their quality from images and micro-climate

data. These insights informed our formulation and predictive modeling of their biological development process, enabling us to optimize fruit-specific harvest plans for short-term demand. We also elevated the analysis to a strategic level, leveraging decision-focused learning (DFL) to forecast productivity (yield) and market (prices), thereby supporting growers in making more profitable selling decisions. Below, these questions are answered specifically according to these findings, with the relevant chapters noted for each.

**I. How accurately can key quality attributes of fruit, such as external characteristics like ripeness and internal properties like sugariness, be estimated using in-field data? How do these estimations compare to traditional assessments or benchmark methods based on laboratory data?**

▷ Chapter 3, Chapter 4, Chapter 5

We proposed two main conceptual methodologies, one for measuring the growth based on (longitudinal) observations, and the other for performing non-destructive quality estimation of internal properties such as sugariness (Brix) and firmness.

The former was built upon revisiting the evaluation criteria” instead of using ML models to imitate human horticulturalists’ evaluations, we considered directly taking the hue from camera observations. This approach provides consistent, objective, and quantitative evaluations of the entire growth cycle, moving beyond subjective human perception. Our comparison to human-based labeling in Section 5.4.1 shows that while overlaps exist between fine-grained categories due to human subjectivity, objective color metrics provide a more reliable measure, supporting their use in a continuous scoring system.

For the latter, our models predict strawberry Brix with competitive accuracy using in-field data, collected from easily-acquired devices. The results confirm the feasibility of reliably anticipating fruit quality pre-harvest, supporting better harvesting and supply strategies. This non-destructive method is suitable for in-the-wild use and has been integrated into a mobile app with offline models to demonstrate its feasibility in real practice.

**II. How important are the image and micro-climate sensor data in training proper (internal) quality prediction models? What is an optimal way of utilizing this multi-modal data?**

▷ Chapter 4

Our analysis reveals the complementary roles of environmental and image data for predicting internal fruit quality. While environmental data – particularly records from the past fourteen days – proved to be more influential, it cannot pinpoint fruit-specific qualities on its own. Image data provides this crucial specificity, yet is insufficient on its own for robust prediction.

This synergy explains why an integrated, multi-modal approach is optimal. This is confirmed by our results, where the optimal models integrated image features with weather-based, aggregated Brix predictions, achieving a significant reduction in RMSE over baseline methods and related works.

**III. What are the distinct characteristics of a long-term monitoring task compared to typical MOT tasks? How will they affect the performance of existing MOT algorithms, compared to their benchmarks on typical MOT tasks?**

▷ Chapter 3

Tracking the growth of biological objects presents a distinct MOT challenge, primarily due to the significant temporal evolution and sparse data collection during a prolonged period. To investigate this, we curated the “*The Growing Strawberries Dataset (GSD)*”, which exhibits two characteristics that differ from typical MOT benchmarks such as *MOT20*: (1) progressive appearance changes driven by biological development, and (2) irregular object movements resulting from a low frame rate.

These factors create a significant challenge for existing MOT algorithms. The slow but drastic appearance changes make it difficult to maintain consistent object associations over time, while the irregular movements increase the risk of identity switches. Our benchmarking of five state-of-the-art MOT algorithms on *GSD* confirms these limitations, revealing a marked performance decline in long-term tracking for such dynamic objects, as detailed in Section 3.4.5.

**iv. Apart from MOT metrics, how can monitoring results be practically assessed? In what ways can they be combined with in-field quality analysis to effectively represent biological development?**

▷ Chapter 3, Chapter 5

For a practical assessment beyond standard MOT metrics, we first quantify ripeness through color hue analysis, a method also central to our answer for Research Question I, and then use these ripeness indicators to model the complete growth trajectory of each fruit. By mapping this maturity data to MOT annotations, we find that the biological development of strawberries follows a typical S-shaped “growth function,” a finding that aligns with established horticultural research. These functions reveal growth trends and rate variations under different cultivation conditions.

Interestingly, this process also highlights a key challenge for many state-of-the-art MOT algorithms: their performance, particularly with increased trajectory fragmentation and ID switching, more often when a fruit’s color changes rapidly (Figures 3.6 and 5.6). This finding not only highlights the significant challenge that appearance changes pose in the MOT field but also underscores the value of incorporating physical properties – such as the color development that we demonstrated – when examining real-world challenges.

**v. How can we analyze and apply individual growth patterns at the fruit level to support the decisions and optimization at the level of a whole greenhouse?**

▷ Chapter 5

Following the answers to Research Question iv, once individual growth functions are established, they can be aggregated to train predictive models that forecast the development of new fruits. In the CROPi framework (Chapter 5), we explored two such modeling approaches: a traditional parametric growth function and a novel “growth language” model inspired by NLP techniques.

**vi. How can individual fruit growth models be used to plan the harvest and optimize the greenhouse’s production?**

▷ Chapter 5

Crop yield estimation can be approached in two ways: as a multivariate time-series analysis or as an extended object detection task, which counts the number of maturing fruits in sample areas and scaling up the results. Our research primarily focuses on the former approach, employing a popular DFL method called “smart predict-then-optimize” (SPO). This method incorporates the downstream impact of predictions into the training of predictive models, enhancing their relevance for decision-making.

In the meantime, we retain the latter approach as a complementary strategy to tailor the final yield by controlling and rescheduling the harvests to some extent. This involves formulating and fitting growth models to observable fruits, a need previously addressed by Research Question iv. The outcomes of these growth models are crucial for optimizing harvest timing and improving overall production management.

**vii. How can the investment performance be improved by making the yield forecasting model aware of the downstream economic impacts of their predictions? What is the effect on the model and on the profitability?**

▷ Chapter 6

Having established methods to monitor, predict, and optimize the physical harvest at the fruit and greenhouse levels, the final strategic challenge is to maximize the financial performance of that supply in the market. To address this, we frame the decision of how to sell the harvested produce as a portfolio optimization problem in Chapter 6. The goal is to maximize profitability by optimally balancing fixed-price contracts with forecasted opportunities in (volatile) real-time markets, all while accounting for an uncertain and not fully controllable production volume. While this scenario was initially developed for our strawberry case, this framework is also applicable to a broader range of real-world situations, such as trading other agricultural products or renewable energy.

We explored two approaches: prediction-focused learning (PFL), which uses prediction accuracy as the loss function, and decision-focused learning (DFL), which uses the SPO loss (“regret”). Our analysis of models trained with the SPO loss revealed a common performance discrepancy: models trained solely on regret did not consistently minimize testing regret. Additionally, we found that combining MSE and SPO losses led to different biases, yet similar layer weights and self-attention coefficients. This effect was relatively more evident in the price prediction model, where the output does not constrain the objective function of the downstream optimization problem.

**viii. What is an effective strategy to predict and realize greenhouse production, so as to achieve the best alignment of supply with downstream parties?**

▷ Chapters 5 and 6

Our ultimate solution is a two-stage method that integrates strategic, long-term commitments with dynamic, short-term harvest planning. The first, strategic stage establishes the overall production goals by defining long-term supply targets, which are often fixed by contract farming. These targets are informed by the improved yield and price forecasting algorithms developed in our answer to Research Question vii. The second, operational stage then uses individual fruit growth models to determine the real-time readiness and availability of the crop, as addressed in our answer to Research Question vi.

This process culminates in a harvest plan where the predicted supply of ready fruit (from the operational stage) is matched against the daily requirements derived from the long-term contracts (from the strategic stage). By determining the ideal harvest date for each fruit, this method ensures that the overall harvesting decisions to be best aligned with the greenhouse's most important strategic commitments, creating a more profitable and practical production plan.

## 7.2 Outlook and Insights

Building on the paper-based chapters, this section synthesizes three overarching insights that integrate and contextualize the key findings: the transformative potential of AI, the pivotal role of data, and the importance of dataset consistency in achieving impactful and scalable solutions. Collectively, they underscore how machine learning transforms operational data into actionable insights, enabling non-destructive and automated quality assessment which, in turn, informs and optimizes greenhouse's decisions, while also suggesting promising directions for future research and development.

### AI Enables Harvest and Portfolio Optimization for Growers.

This thesis focused on how AI can enhance the daily operations of small growers, who rely on supply contracts for stability about what will be sold and when, yet face challenges from limited negotiating power and the inherent uncertainty of agriculture. To navigate this uncertainty, growers employ several operational strategies: (i) utilizing more flexible but less stable selling channels such as local markets; (ii) modifying the harvest schedule by advancing or delaying harvest times; and (iii) employing buffers like cold storage to manage surplus. Effectively balancing the stability of long-term contracts with these short-term adjustments is the most critical challenge. Therefore, our ultimate solution is a two-stage method that integrates strategic, long-term commitments with dynamic, short-term harvest planning. The strategic stage, addressed in Chapter 6, focuses on optimizing contract portfolios across multiple sales channels. The operational stage, detailed in Chapter 5, then uses these strategic goals to guide daily harvest decisions.

For the strategic stage, our work in Chapter 6 explore how ML models can balance competing demands from multiple selling channels. Our findings highlight that DFL essentially tailors bias layers in neural networks to steer decision-making toward risk aversion. This effect is particularly pronounced when the model output serves as a decision parameter but not constraining the problem. However, DFL is not a one-size-fits-all solution. When decisions rely on multiple model outputs, we recommend initializing training with independent loss functions, such as mean squared error (MSE), to capture essential feature patterns efficiently as the first step.

For the operational stage, we introduce CROP1, an end-to-end framework for optimizing harvest schedules. In Chapter 5, we frame the task as an assignment problem, aiming to maximize the alignment between predicted fruit quality and demand at specific time slots. While the demonstration uses color as a proxy for quality, the framework is extensible to other metrics such as size and firmness. A key limitation of the current setting

is the assumption of risk-neutrality, hence, making stochastic, robust, or scenario-based optimization methods more practical alternatives for risk-averse decision-makers.

Looking ahead, future work could incorporate *storage* as the third optimization strategy. For instance, it could address a key limitation in the CROPi framework by offering a more realistic alternative to its current handling of surplus fruit, which simply assumes all over-productions can be sold at a fixed discount/cost. Truly integrating storage, however, requires moving beyond the fixed-horizon Hungarian method used for daily assignments. This extension of the problem would necessitate advanced techniques such as dynamic programming, Monte Carlo simulations, or reinforcement learning to optimize longer-term decision-making. Moreover, adopting such methods would also require a more complex market dynamics model from Section 6.5. Tackling these emergent challenges can be the next frontier for this research, promising a truly unified AI solution for building more resilient and profitable grower portfolios.

## Analytical Methods Support Objective Fruit Assessments.

In the market, fruit is categorized by attributes like ripeness, size, shape, Brix, firmness, shelf life, and etc., which can enable price discrimination [123, 195]. However, this reliance on subjective grower expertise frequently leads to disagreements with downstream customers over quality standards, resulting in significant financial losses, food waste, and supply chain inefficiencies. Our research addresses this by developing objective, data-driven evaluation methods, whose practical adoption is increasingly viable due to the rapid advancement of fast, portable computer vision (CV) models.

### 7

Ripeness assessment is a prime example, which is often subjective because reference standards are not always consistently applied in real practice. We address this by proposing raw color hue as a direct, quantitative indicator of ripeness, rather than training ML models to mimic flawed human evaluations as most related works did. This hue-based approach was first validated against human perception in Chapter 5 and then extended into a parametric growth function. The resulting growth model proved versatile, as we also used it in Chapter 3 to visualize and diagnose failure modes of MOT algorithms. This progression demonstrates how a simple, objective metric can be foundational for developing powerful analytical tools with applications far beyond initial quality assessment.

Another challenge is destructive testing, required for metrics like Brix or firmness. Achieving a reliable estimate requires sacrificing a large proportion of harvests, which causes waste. To address this, Chapter 4 details a group of ML- and CV-based methods to predict Brix using multi-modal data. Our models demonstrate an accuracy comparable to lab-based benchmarks on in-field data, showcasing their better efficacy and practical viability. Because the models were trained on the growth monitoring dataset *GSD*, future work can leverage early-stage images to forecast Brix or even its development.

Building on these proofs of concept, our data-driven approach also holds the potential to be extended to other key attributes. For instance, CV offers a promising, non-invasive alternative for size and shape measurement, though robust models require larger datasets to tackle the problem of partial visibility. Firmness, another destructive test, can likely be addressed with methods similar to our Brix prediction [123]. Shelf life, however, presents



a more complex challenge, as our current models cannot be directly transferred to this task. Overall, synthesizing these methods can provide the supply chain with a unified, data-driven language for assessing fruit quality in a non-destructive manner, ultimately enhancing the efficiency and profitability.

## **Dataset Consistency and Sufficiency are Critical for Solutions when Spanning Multiple Seasons.**

This project produced two complementary datasets using affordable, real-world hardware: a time-lapse image dataset (“the Growing Strawberries”/*GSD*) for longitudinal growth monitoring, and a multi-module dataset linking images to detailed quality assessments. These resources enabled our research at the fruit and greenhouse levels, while market-level analysis utilized public data.

*GSD*, created to document long-term biological development, presents unique challenges that are actually common in real-world data collection. When forming an MOT task, the fruits’ irregular movement and evolving appearance create conditions rarely encountered in standard MOT datasets. Furthermore, because the data spans multiple daytime and seasons, it highlighted the critical importance of data consistency. We found that subtle changes in field illumination caused significant fluctuations in color hue, which complicated the evaluation in Section 5.4.1 by causing measurements to misalign with human perception. When physical interventions during collection are infeasible, robust post-processing becomes essential. While algorithmic techniques like color correction offer limited mitigation, deep learning methods such as generative models achieve higher output consistency, yet with the risk of altering the actual color. This trade-off between algorithmic performance and data fidelity calls for future work on developing advanced models that can ensure color consistency without distorting the underlying data.

The multi-module dataset combines images from *GSD* with critical cultivation data like micro-climate records, to enable research on non-destructive quality prediction and precision agriculture. However, its sufficiency for training large neural networks is a key limitation. While techniques like transfer learning can maximize the utility of the existing data, developing truly robust models still requires more extensive datasets. This limitation underscores the need for greater collaboration and a commitment to open science. By collectively expanding and sharing foundational resources, the research community can accelerate progress and foster a more impactful research environment.





# Appendix

## A.1 Hosting and licensing information

### License

*GSD* is released under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

### Terms of Use

By accessing and using *GSD*, users agree to comply with the terms and conditions outlined in the CC BY-NC-ND 4.0 license. Users are responsible for ensuring the appropriate use of the dataset in accordance with the license and any applicable laws or regulations.

### Author statement

The corresponding authors state that they collected the data as described in this document and in the main paper. The authors have the right to publish this dataset. *GSD* is licensed under the CC BY-NC-ND 4.0 license. Users of this dataset are required to comply with the license terms, including providing proper attribution when using the dataset. We provide the dataset “as is”, without any warranty or guarantee of its accuracy or reliability. We disclaim any liability for errors, damages, or consequences arising from the use of the dataset.

### Hosting and Maintenance Plan

*GSD* is hosted and maintained on *4TU.ResearchData* Platform. It is published with a DOI [doi.org/10.4121/e3b31ece-cc88-4638-be10-8ccdd4c5f2f7.v2](https://doi.org/10.4121/e3b31ece-cc88-4638-be10-8ccdd4c5f2f7.v2) for long-term accessibility and versioning under the CC BY-NC-ND 4.0 license.

The *extended-GSD* and *dark-GSD* is hosted and maintained on *4TU.ResearchData* Platform. They are published under [268] and [267] for long-term accessibility and versioning under the CC BY-NC-ND 4.0 license. Overall, we made a collection of all the growth monitoring images at [doi.org/10.4121/f2ad72ce-3e5b-429e-ba14-78eed903ce03](https://doi.org/10.4121/f2ad72ce-3e5b-429e-ba14-78eed903ce03).

The measurement and environment dataset is hosted and maintained on *4TU.ResearchData* Platform. The data accompanying Chapter 4 is published at [doi.org/10.4121/21864590.v1](https://doi.org/10.4121/21864590.v1). The collection of all the measurement and environment data over the four years, as presented in Chapter 2 is published at: [doi.org/10.4121/1d02156d-7011-4052-8175-da52a2e32cba](https://doi.org/10.4121/1d02156d-7011-4052-8175-da52a2e32cba) for long-term accessibility and versioning under the CC BY-NC-ND 4.0 license.

The code that we used in Chapter 6 are available at: [anonymous.4open.science/r/predict-then-contract-F6EB/](https://anonymous.4open.science/r/predict-then-contract-F6EB/). The GitHub address will be available with the camera-ready version.

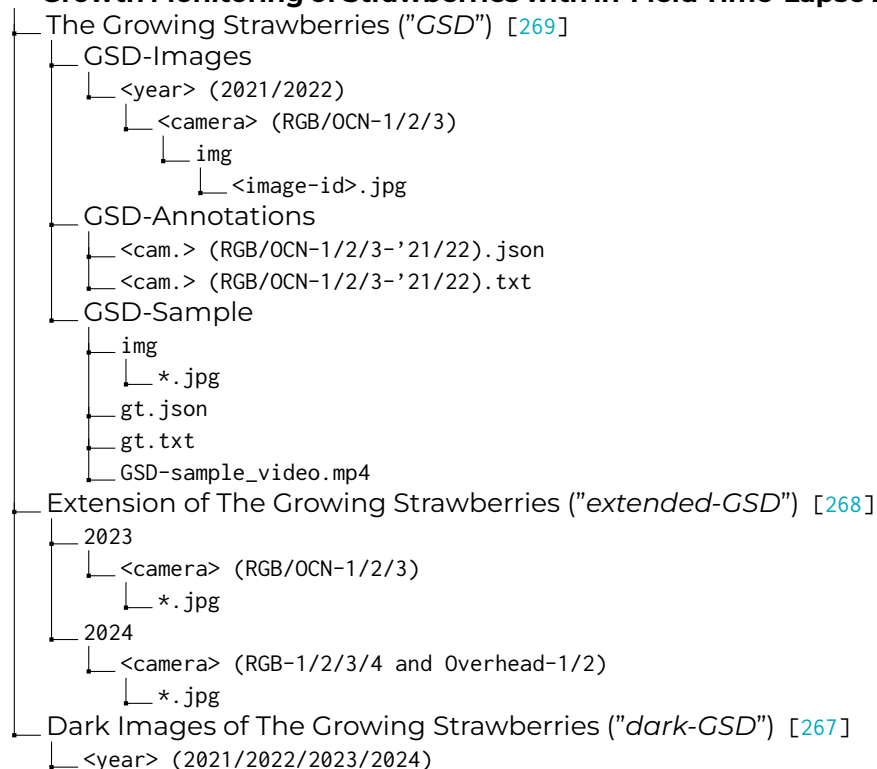
## Dataset Structure

To suit studies in different fields, we separate our datasets into two main sets: the *time-lapse images* for growth monitoring, and the relevant data to conduct *quality prediction* studies. This section presents the abstracted structure and the connection of the two sets. Detailed licensing and hosting information can be referred to Appendix A.1.

### Time-lapse images

Part of the *time-lapse images* set, collected in 2021 and 2022, has been annotated and published with the paper Wen et al. [272], it is named as *The Growing Strawberries Dataset (GSD)* and hosted by the *4TU.ResearchData* Platform: Wen et al. [269]. Along with the main MOT dataset, we provide a sample subset of *GSD* under the file folder *GSD-Sample*. The sample includes images collected from 2021-09-01 to 2021-09-02 by *RGB-3*, and the corresponding annotations in a *coco-format* JSON file and a *TXT* file compatible with the MOT evaluation tools. The images are with the original filenames assigned by the cameras when the photos were taken. We also provide a short video to illustrate a subsequence of *GSD*. The video presents the growth monitoring of strawberries from 2021-09-01 to 2021-09-07 in *RGB-3*. The non-annotated images collected in 2023 and 2024 are archived on *4TU.ResearchData* Platform, of which the daytime images are established as "*extended-GSD*" and the darker images are stored together with "*dark-GSD*". Overall, we made a collection of all the growth monitoring images at Wen et al. [270].

### Growth Monitoring of Strawberries with in-Field Time-Lapse Images



```

├── <cameras of each year>
│   ├── *.jpg
│   └── annotations (for 2021/2022)
│       └── <cam.> (RGB/OCN-1/2/3-'21/22).json

```

## Quality prediction

The subset of the *quality prediction* dataset collected in 2021 was published as the accompanying dataset of Wen, Abeel, and de Weerd [\[264\]](#), which benchmarked the Brix prediction methods and is also the Chapter 4 of this thesis. The prediction about ripeness and firmness using convolutional neural networks (CNN) was analyzed in Jol, Wen, and Van Gemert [\[123\]](#). Since the cultivation performance can vary from year to year, we still keep collecting and extending the dataset over more seasons. For instance, Appendix A.7 include new measurement results from different greenhouses and cultivation years to validate the models demonstrated by Wen, Abeel, and de Weerd [\[264\]](#). To include different perspectives of quality measurement labels with the in-field images collected at different years, we re-organized the entire dataset and archived them as an entire group in the *4TU.ResearchData* Platform, while the original dataset was still available in the same platform [\[266\]](#).

### Strawberry Quality Prediction with in-Field Data

```

├── Individual quality evaluations
│   ├── RGB segments
│   │   ├── <year>_<day-of-year>_<fruit-id>.png
│   │   └── <year>_<day-of-year>_<fruit-id>_bbox.png
│   ├── OCN segments
│   │   ├── <year>_<day-of-year>_<fruit-id>.png
│   │   └── <year>_<day-of-year>_<fruit-id>_bbox.png
│   ├── Measurement and labeling
│   │   ├── Measurements_with_Image_Connections-<year>.csv (2021/2022/2024)
│   │   └── On-shelf images (2022)
│   │       └── *.jpg
│   └── Micro-Climate
│       └── Greenhouse_Environment-<year>.csv (2021/2024)
└── Aggregated quality evaluations
    └── Strawberry_Measurements-<year>.csv (2021/2022/2023/2024)

```

The original dataset [\[266\]](#) was structured as in below. The image part of the dataset contains the growth-monitoring images with their original names, which has the corresponding timestamp associated. No specific format is considered.

Data underlying the research of quality prediction of strawberries with RGB image segments [\[266\]](#)

```

├── Segments
│   └── <date>_<fruit-id>.png
├── Images
│   └── *.JPG
├── Strawberry_Measurements_with_Seg_Connections_mtd1.csv
├── Greenhouse_Environment_Hourly_20210401-1118.csv
└── Strawberry_Plant_Load_2021.xlsx

```

## A.2 Implementation details for Chapter 2

This section introduces the detailed data collection hardware setup of GSD.

### Data collection setup

The strawberries in the greenhouse were cultivated in planting baskets, which were hung in parallel lines. Figure A1 gives a side view of the rows. Strawberries grew out from both sides of the baskets.



Figure A1: A side view of the planting baskets. The cameras were attached to the heating pipe at the neighbor's row. For example, if the strawberries grew in the left row in the image, cameras would be installed at the highlighted heating pipe. This particular image is not taken by the data collection devices, so the distortion in the image is not related to the strawberry observations.

The cameras were grouped as three pairs of RGB and OCN cameras. They were installed on the opposite row from where the strawberries were growing, as shown in Figure 2 in the main text. As Figure A2 shows, they were fixed on the heating pipe with camera clamps. They were connected to the local electrical grid with a powered USB hub, so they could stay awake all the time.

### Camera Settings

We used cameras from MAPIR® to collect the time-lapse images. We used *Survey3N - Visible Light RGB* for RGB image collection and *Survey3N Camera - Orange+Cyan+NIR (OCN, NDVI)* for OCN images. All cameras had the same settings as we show in Table A1. We set the shutter speed, ISO, and white balance into the auto to maintain an optimal state of each image. No extra exposure was added. We take the neutral setting of color presentation, contrast, and sharpness.



Figure A2: Camera installation for data collection. The cameras were fixed to the heating pipe and connected to the electrical grid with the yellow USB hub.

Table A1: Settings of the data collection cameras. Hereby all the changeable settings of the cameras are listed. There is no difference when setting the RGB or the OCN cameras.

Item	Setting
Shutter	Auto
ISO	Auto
White Balance	Auto
Exposure	0
Metering	Centre
Color	Normal
Contrast	Medium
Sharpness	Medium

## A.3 Implementation details for Chapter 3

We introduce further details about the implementations of the algorithms and explain the reason of our parameter settings in this section.

All the object detection and MOT algorithms that we used for the benchmarking experiments are open-source: YOLOX-x and Faster R-CNN (model established with the Detectron2 framework [278]) come with Apache License 2.0. ByteTrack and OC-SORT use the MIT License. DeepSORT and StrongSORT use GNU General Public License v3.0.

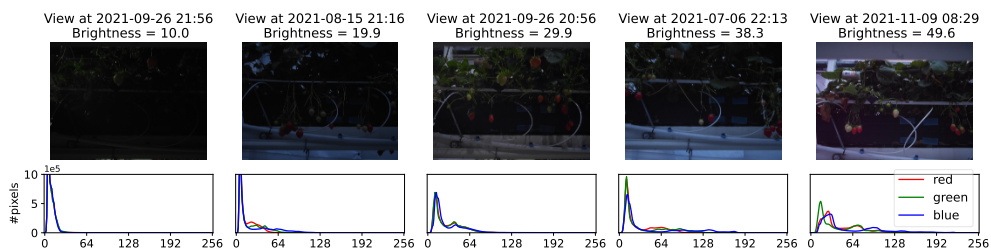


Figure A3: Example dark images at different brightness levels. The collection time and the average Luma are written on top of each image. The RGB spectrum is drawn beneath.

## Selection criteria of daytime subset of image data

We have divided the set of images into a daytime subset, which has a brightness (Luma) of at least 50, and a dark image set which is not annotated. We calculated the brightness (Luma) of images according to [29], and here exemplify the brightness levels in figure A3. We illustrate the example images and the corresponding Luma to show that 50 is a rational threshold to select the “day-image” and “dark-image” subsets. An overview of the proportions of images with different levels of brightness is shown in figure A4.

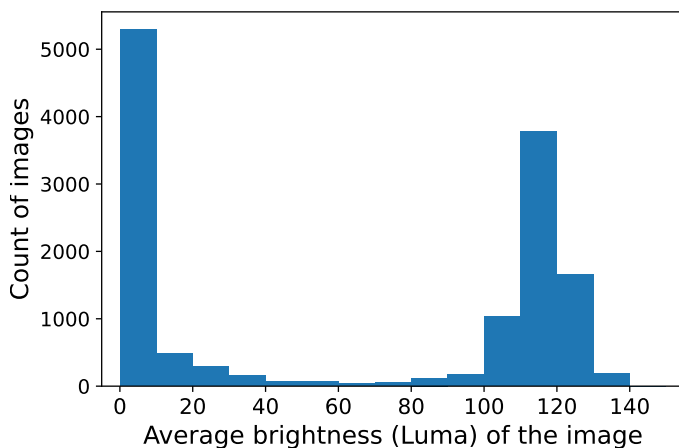


Figure A4: Histogram of GSD images under different ranges of brightness. The brightness value is calculated in Luma and averaged over all pixels.

## Detection models

We trained two object detection models, YOLOX-x [87] and Faster R-CNN [207], to perform the object detection stage of the MOT algorithms. Since they had similar performances, as presented in appendix A.6, we selected the MOT results produced only from the YOLOX-x model’s predictions in the main paper.

This section presents the hyper-parameters for building up and training the models. Both models are trained with the complete RGB image dataset, with a “leave-one-camera-out”



train-test-split strategy since the RGB cameras monitored different plants. The two models shared the same scales of data augmentation, listed in table A2. The detection results are filtered with a confidence threshold of 0.1 before going to the association stages of the MOT algorithms.

Table A2: Data augmentation for training the object detectors. The first column lists how we augment the data, and the second column indicates the value ranges.

Data Augmentation Method	Scale
Random flip horizontal	(probability =) 50%
Random flip vertical	(probability =) 50%
Random rotation	0-90 degrees
Random brightness	$\times 0.92-1.12$
Random contrast	$\times 0.92-1.12$

The YOLOX-x model was initialized with a COCO[152] pre-trained model. The images are scaled to 2133 $\times$ 1600 and then padded to 2174 $\times$ 1600 to fit the input aspect ratio. The batch size for training is 4. The model is trained with a cosine annealing learning rate  $3.125e^{-6}$  with a warm-up, and a weight decay of  $5e^{-4}$ . The model is trained by 100 epochs on an Nvidia Tesla V100 GPU. We select the checkpoint with the optimal parameters on the validation set to predict the object detection results for further steps of the experiments.

We trained another Faster R-CNN model using the original ResNet-50 from MSRA [108] and Feature Pyramid Network (FPN) [151] as the model backbone. The model is pre-trained with ImageNet [68].

## MOT algorithms

Before starting the evaluations of the MOT algorithms, we first conducted grid searches to figure out the optimal parameters for the strawberry growth-tracking scenario. The grid search was conducted on the YOLOX-x detections of the *RGB-1* set.

table A3 and A4 present part of our grid-search results. As is shown, the Intersection over Union (IoU) threshold (“iou-thre”) was the dominant variable of performance of OC-SORT, both in terms of MOTA and IDF1. One reason could be the irregular movements of objects, illustrated by the Figure 4 in the main text. The confidence threshold had limited effects when using low IoU Threshold.

DeepSORT uses the maximum cosine distance of features (“max-cos-dist”) as a gating threshold. Considering the changing appearance of the *GSD* objects, we checked the cosine distance of the features of the same object over the frame. As depicted by figure A5, the features of adjacent observations of the object have an average cosine distance of 0.44. Hence, we regard distances larger than the value are large enough for distinctive objects. Therefore, we select 0.45 as the max-cosine-distance value when implementing DeepSORT and StrongSORT.

The final decision on the parameters is made by referring to the grid search results and the default settings of the MOT algorithms. Details are listed in table A5.



Table A3: Grid search of iou-thre and conf-thre in OC-SORT. Performances are indicated by MOTA. All experiments have a default setting of min-hits=3 and max-age=30. The selections of conf-thre are the indices of rows, and the selections of iou-thre are indicated by the columns.

Confidence Threshold	IoU Threshold				
	<b>0.1</b>	0.3	0.5	0.7	0.9
<b>0.1</b>	64.5	61.4	56.1	45.6	15.7
0.3	64.6	61.4	56.2	45.6	15.7
0.5	64.5	61.4	56.1	45.6	15.7

Table A4: Grid search of iou-thre and conf-thre in OC-SORT with performances indicated by the IDF1 score under different settings. All the experiment settings are the same as in Table 2. The selections of conf-thre and iou-thre are the indices of rows and columns respectively.

Confidence Threshold	IoU Threshold				
	<b>0.1</b>	0.3	0.5	0.7	0.9
<b>0.1</b>	67.3	64.9	60.6	51.2	20.2
0.3	67.4	64.9	60.6	51.2	20.2
0.5	67.3	64.8	60.5	51.2	20.2

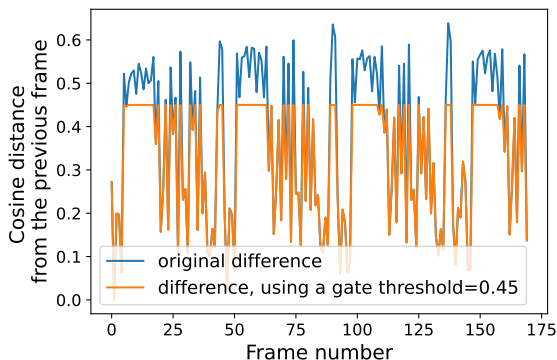


Figure A5: The cosine distances of the normalized features of adjacent observations during the complete track of an example strawberry. The x-axis shows the temporal sequence of the observations, and the y-axis indicates the cosine distance from the previous observation. The features are encoded by the same extractor as we used for DeepSORT and StrongSORT. The blue line presents the original cosines distances, and the orange line depicts the effect of adding a gate threshold = 0.45.

Table A5: Detailed parameter settings of the benchmark experiments of the MOT algorithms.

	conf. thre	iou thre	min hit	max age	max cosine-cost
OC-SORT	0.1	0.1	3	30	-
ByteTrack	0.1	-	3	30	-
Deep-SORT	0.1	0.1	1	30	0.45
Strong-SORT	0.1	0.1	1	30	0.45

## A.4 Implementation details for Chapter 4

### Time span

We archived hourly environmental records and weekly plant load data during the data collection period, as previewed in Table 2.1. We grouped the environmental information as rolling averages every 7 days for the model training.

The plant load was recorded by calendar weeks, from the 12<sup>th</sup> week to the 42<sup>th</sup> week of 2021. As no finer time information was available, the significance of plant load was analyzed in different branches of experiments when all the other environment information was also aggregated by calendar weeks.

### Neural network architectures

We trained two/three/four-layer CNNs from scratch for supervised learning (SL), with kernel size 12\*12, 9\*9, or 6\*6. Each was connected with a two-layer multi-layer perception (MLP) which outputs the Brix value of each strawberry. When considering transfer learning for the SL, we replaced the CNNs trained from scratch with popular pre-trained models such as the *ResNet* and the *EfficientNet*. We used three-layer and four-layer CNNs to build the encoders and decoders for semi-supervised learning (SSL). We considered max-pooling layers of 2\*2, 4\*4 and 5\*5 among the convolutional layers to reduce the volume of the latent space. The MLPs that mapped the latent space to the final output, i.e. the Brix values, consisted of three or four fully connected layers.

### Parameter selection of regression models

We tested four scales of regularizers (also named “alpha” in mathematical models) in the kernelized ridge regression (KRR), with three degrees of the polynomial kernel (“degree”). As is shown by Figure A6, the levels of fitness were illustrated by plotting the predicted Brix value with the first principal component of all the input features, which is all available environment records up to 21 days in advance. We finally decided to use both alpha=1 and alpha=10 in our experiment series.

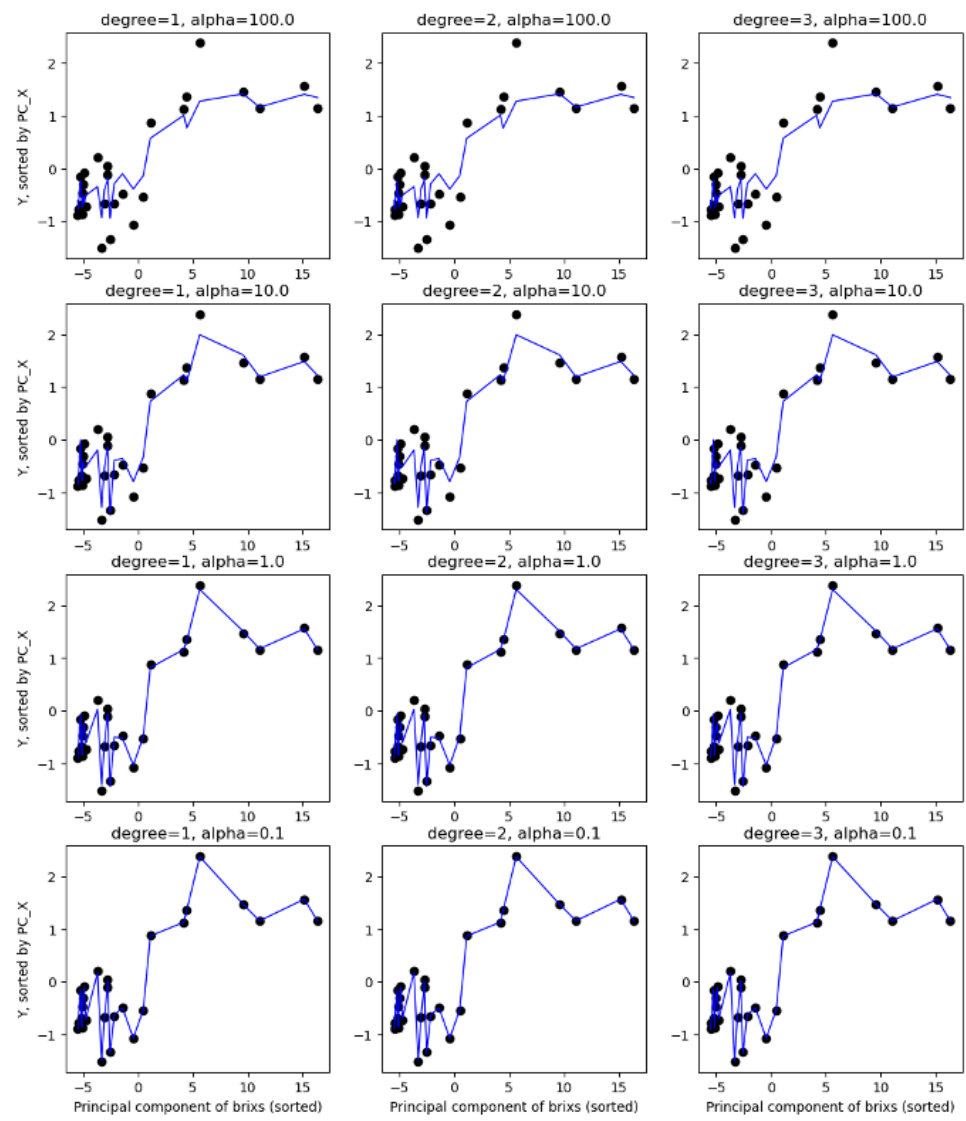


Figure A6: Parameter tests of the regularizer ( $\alpha$ ) and polynomial kernel (degree) in the kernelized ridge regression (KRR). We tested  $\alpha=0.1, 1, 10, 100$  and polynomial degree=1,2,3 to decide the final parameters. The influences from various  $\alpha$  are discussed over rows, and the fitness under different polynomial degrees is compared within each row. The x-axis of each subplot indicates the value of the first principal component of the input feature. The y-axis of each subplot indicates the value of Brix. All the data points are sorted by the scale of the first principal component, so as to better present the influence from the regularizer.

## A.5 Implementation details for Chapter 6

### Model parameters

Table A6 lists all the notations that we use to describe the models in Sections 5.2 and 5.3 of the main text. In the main text, every decision parameter or variable is introduced with a boldface, and used in *italic* in later texts.

Table A6: Annotations of decision parameters and variables.

Notation	Description
$Pft$	<b>Objective function:</b> the profit. It is a function of $Y, P_c, P_r, \mathcal{D}_c$ .
$\mathcal{D}_c$	<b>Decision variable:</b> the amount of product that is planned to be sold by the contract.
$\tilde{\mathcal{D}}_c$	The decision of contract transaction made under predicted parameters.
$t$	The time gap between when a contract is signed and when the transaction is executed.
$Y$	The production/yield of the product, in an artificial volume unit.
$P_c$	The unit price of the product sold by the <i>contract</i> , in an artificial monetary unit.
$P_r$	The unit price of the product sold at the <i>market</i> .
$\hat{Y}, \hat{P}_r$	The predicted values of the production $Y$ and market price $P_r$ .
$(a_{ct} \cdot P_c + b_{ct})$	The maximum transaction capacity of the contract, also written as $a_{ct} \cdot P_c + b_{ct}$ using the fully competitive market model.
$(a_{rt} \cdot P_r + b_{rt})$	The maximum transaction capacity of the market, also written as $a_{rt} \cdot P_r + b_{rt}$ using the fully competitive market model.
$T_c$	The actual amount of product sold by the contract at the transaction date, limited by the pre-made decision, the actual production, and the contract capacity, i.e. $T_c = \min\{\tilde{\mathcal{D}}_c, Y, T_{c,\max}\}$ .
$T_r$	The actual amount of product sold at the market at the transaction date, limited by the unsold items from the contract and the market capacity, i.e. $T_r = \min\{Y - T_c, T_{r,\max}\}$ .
$a_{c1}$	The unit cost of the production.
$a_{c2}$	The unit cost of delivery shortage of the contract.
$a_{c3}$	The unit cost of over production waste.

### Situation-wise object function

Due to the time gap  $t$  between committing and executing transactions to the contract, the contracting decision can only be made under a few parameter predictions, i.e.  $\tilde{\mathcal{D}}_c = \arg\max_{\mathcal{D}_c} Pft(\mathcal{D}_c, P_c, \hat{P}_r, \hat{Y})$ . Due to the limitation from contract capacities and production amount, there can be varied consequences: For decisions within the feasible region, i.e.  $\mathcal{D}_c \geq 0, \mathcal{D}_c \leq Y, \mathcal{D}_c \leq (a_{ct} \cdot P_c + b_{ct})$ , we have  $T_c = \min\{\mathcal{D}_c, Y, (a_{ct} \cdot P_c + b_{ct})\} = \mathcal{D}_c$ . Since  $P_c$  is known when contracting, the only constraint that the decision can violate is when  $\hat{Y} > Y$

results in  $\tilde{\mathcal{D}}_c > Y$ . In that case, we have  $T_c = \min\{\tilde{\mathcal{D}}_c, Y\} = Y < \tilde{\mathcal{D}}_c = \mathcal{D}_c \leq \min\{\hat{Y}, (a_{ct} \cdot P_c + b_{ct})\}$  and we pay the supply shortage fine:  $a_{c2} \cdot (\tilde{\mathcal{D}}_c - Y) = a_{c2} \cdot (\mathcal{D}_c - Y)$ , which acts as the penalty of the constraint set. In short, we can re-write equation (2) in the main text into a situation-wise form:

$$\begin{aligned} & \text{Pft}(\tilde{\mathcal{D}}_c, Y, P_c, P_r) \\ &= P_c \cdot T_c + P_r \cdot T_r - a_{c1} \cdot Y - a_{c2} \cdot \text{pos}[\tilde{\mathcal{D}}_c - Y] - a_{c3} \cdot \text{pos}[Y - T_c - T_r] \\ & \text{when } Y - \tilde{\mathcal{D}}_c < 0 : \\ &= P_c \cdot Y - a_{c2} \cdot (\tilde{\mathcal{D}}_c - Y) \\ &= -a_{c2} \cdot \tilde{\mathcal{D}}_c + f_1(Y, P_c) \end{aligned} \quad (1)$$

$$\begin{aligned} & \text{when } 0 \leq Y - \tilde{\mathcal{D}}_c \leq (a_{rt} \cdot P_r + b_{rt}) : \\ &= P_c \cdot \tilde{\mathcal{D}}_c + P_r \cdot (Y - \tilde{\mathcal{D}}_c) - a_{c1} \cdot Y \\ &= (P_c - P_r) \cdot \tilde{\mathcal{D}}_c + f_2(Y, P_r) \end{aligned} \quad (2)$$

$$\begin{aligned} & \text{when } 0 \leq Y - \tilde{\mathcal{D}}_c > (a_{rt} \cdot P_r + b_{rt}) : \\ &= P_c \cdot \tilde{\mathcal{D}}_c + P_r \cdot (a_{rt} \cdot P_r + b_{rt}) - a_{c1} \cdot Y - a_{c3} \cdot (Y - \tilde{\mathcal{D}}_c - (a_{rt} \cdot P_r + b_{rt})) \\ &= (P_c - a_{c3}) \cdot \tilde{\mathcal{D}}_c + f_3(Y, P_r) \end{aligned} \quad (3)$$

Here,  $f_*(\dots)$  denotes a function that depends solely on the parameters within the parentheses (bracelet) and is not influenced by external factors, such as  $\tilde{\mathcal{D}}_c$ . As shown above,  $\text{Pft}(\tilde{\mathcal{D}}_c, Y, P_c, P_r)$  is a piece-wise linear function of  $\tilde{\mathcal{D}}_c$ . Hence, its second-order derivative is not constantly equal to 0, and SPO loss is feasible to implement in model training.

## Illustration of problem setting

Figure A7 illustrates the training setting of the models and how the losses, which can be either the regret (SPO loss) or the mean-squared error (MSE), are back-propagated.

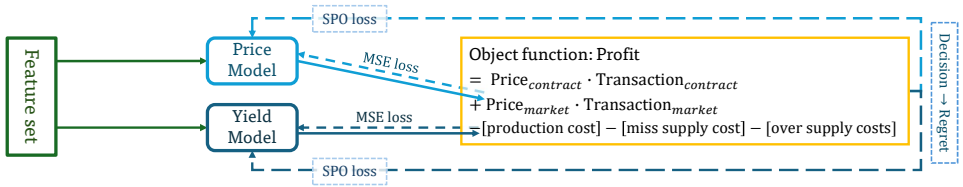


Figure A7: The training setting.

## Case-specific configuration

### Wind farm case

We set up the profit function of the wind farm case using the parameter values as in Table A7<sup>1</sup>. Among all data points, the contract's capability bound 12% of the optimal decisions, and the actual yield bounds 9% of the optimal decisions. Moreover, 6% of the optimal decisions are valued at 0, when the real-time price is higher than the day-ahead price and the market capacity is also sufficient.

Table A7: Parameters to calculate the profit in the wind farm case.

Name	Value	Note
$a_c$	-0.269	Weight from fitting the mapping of day-ahead price to a proportion* of total energy load by ridge regression (RR, alpha=10).
$b_c$	0.774	RR bias from fitting the mapping by plus a small intercept, such that the yield exceeds the contract capacity in ca. 1/3 of the cases.
$a_r$	-0.454	RR weight from fitting the mapping of real-time price to the proportional load.
$b_r$	0.613	RR bias plus a small intercept, such that the yield exceeds the total capacity in less than 10% of the cases.
$a_{c1}$	0.005	A smaller number than that of the tomato case.
$a_{c2}$	0.578	Averaged real-time price.
$a_{c3}$	1.734	3x averaged real-time price.

### Tomato grower case

Using the values from Table A8, the contract's capability bounds 34% of the optimal decisions, and the actual yield bounds 68% of the optimal decisions in the tomato case.

Table A8: Parameters to calculate the profit in the tomato case.

Name	Value	Note
$a_c$	-0.113	RR weight from mapping the contract price to the yield.
$b_c$	2.087	RR bias plus a small intercept, such that the yield exceeds the capacity of the contracts in ca. 1/3 of the cases.
$a_r$	-0.152	RR weight from mapping the contract market price to the yield.
$b_r$	0.391	RR bias plus a small intercept, such that the yield exceeds the capacity of the entire multiple-market system in less than 10% of the cases.
$a_{c1}$	0.010	A larger number than that of the wind farm case.
$a_{c2}$	4.710	3x Averaged market price.
$a_{c3}$	1.570	Averaged market price.

<sup>1</sup>The proportion is calculated as total wind energy generation:total energy generation.

## A.6 Further experimental results of Chapter 3

With the “leave-one-camera-out” policy, we validate the two object detection models in three datasets, i.e. *RGB-1/2/3*. As such, the same policy is also applied to the implementation of the MOT algorithms. The model performances in Chapter 3 are presented by grouping all the validation results together. This section provides the specific performances on each validation set.

### Performance of online MOT algorithms

In the main paper, we present the benchmark of the four MOT algorithms on *GSD* by the overall metrics from the daytime subset. This section shows metrics in specific. Below, table A9 and table A10 present the detailed performance of ByteTrack and StrongSORT respectively. The performance is assessed on each camera subset on the full annotated dataset and on the daytime subset. In general, there are no significant performance gaps in the algorithms when using the data from different cameras. In terms of *RGB-1*, which we used as a test set to decide whether to use a daytime subset or not, we could notice a slight improvement in the MOT metrics, yet the performance on the entire dataset shows limited differences when using different subsets for evaluation.

Table A9: Detailed performance of ByteTrack

Camera	HOTA	MOTA	IDF1	AssA	AssRe	AssPr	IDS/Tr	FM/Tr
PERFORMANCE ON THE FULL DATASET:								
RGB-1	39.77	64.68	40.58	27.30	32.44	66.95	4.4	7.1
RGB-2	39.25	64.65	40.07	27.22	30.82	70.12	5.0	6.8
RGB-3	40.17	80.78	38.09	23.01	25.72	72.64	6.3	7.7
All	39.74	70.29	39.59	25.72	29.49	70.06	5.2	7.2
PERFORMANCE ON THE DAYTIME SUBSET:								
RGB-1	40.03	65.53	40.74	27.38	32.43	66.79	4.4	4.6
RGB-2	39.25	64.67	39.95	27.15	30.64	70.37	5.1	5.8
RGB-3	39.93	81.17	37.40	22.63	25.27	72.49	6.3	5.8
All	39.75	70.73	39.38	25.58	29.26	70.01	5.2	5.4

### Performance of end-to-end MOT algorithms

Since *GSD* consists of long series of high-resolution images, we excluded offline MOT solvers in the scope of benchmarking experiments. However, we argue that end-to-end is feasible for the task, though it does not outperform the other real-time MOT algorithms that we have applied in the paper. We implemented a demo of *GMTracker* [106] on YOLO-X detections of the first few frames of *GSD-2021-CAM-1*. However, the performance metrics are not very positive: without using the quadratic matching function to represent the second-order relationship, the HOTA score is 30 (on the first 1000 frames); and when the quadratic matching function is activated and its GNN uses the parameters trained on MOT17, the HOTA score improved to 38 (on the first 750 frames, as shown in Table 3.2).

Table A10: Detailed performance of StrongSORT

Camera	HOTA	MOTA	IDF1	AssA	AssRe	AssPr	IDS/Tr	FM/Tr
PERFORMANCE ON THE FULL DATASET:								
RGB-1	34.11	33.96	32.93	23.98	28.76	60.36	8.2	6.5
RGB-2	35.31	40.84	33.95	25.34	28.61	65.50	9.1	7.0
RGB-3	37.29	66.36	33.41	21.14	24.01	66.73	11.2	7.3
All	35.51	47.45	33.41	23.38	26.98	64.39	9.5	6.9
PERFORMANCE ON THE DAYTIME SUBSET:								
RGB-1	35.17	36.86	33.83	24.93	30.16	60.94	7.3	4.3
RGB-2	35.76	41.96	34.64	25.89	29.46	65.01	8.5	5.8
RGB-3	37.58	67.74	33.48	21.18	23.95	67.53	10.7	5.4
All	36.14	49.32	33.98	23.87	27.66	64.74	8.8	5.1

We noticed that with the involvement of more frames, the performance slightly raised, nevertheless, the processing time becomes exponentially longer with the greater amount of detections-to-be-matched between frames, as shown in figure A8 – hence we only applied the demo in the first 750 frames of GMT, which has already taken nearly a week to run with a trained object-matching model.

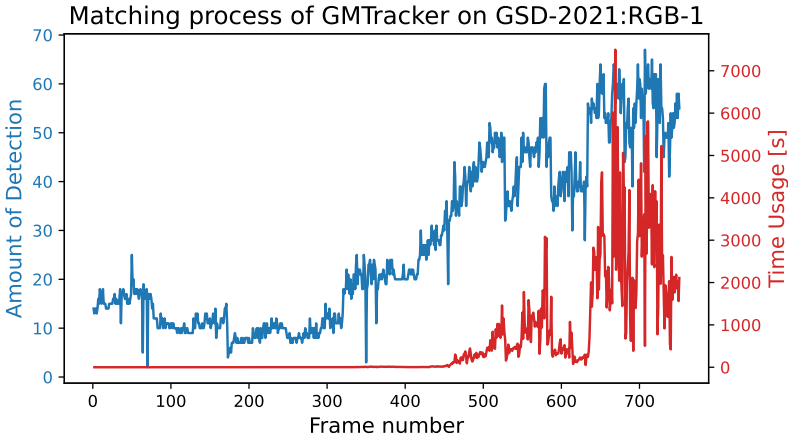


Figure A8: Reaction time of GMTracker when processing each frame of the daytime subset of *GSD-2021-RGB-1*. The x-axis indicates the frame number, which was re-indexed according to the daytime subset. The blue line with the y-axis on the left indicates the amount of detected bbox in each frame. The red line on the right illustrate the time that the model needed to match the detected bbox within that frame with those in the previous frame.

Noted that our implementation used a model with pre-trained parameters, due to the fact that the training process requires tremendous computation effort. Particularly, a quadratic affinity matrix as described in [106] requires much larger memories (e.g. matching 50 objects with another 50 requires 40+GB of memory when training on the *MOT17* dataset)



than the Hungarian-algorithm-based methods, yet could not result in a significant performance increase in our demonstration.

Hence, we did not apply the end-to-end MOT on the entire daytime subset as the other four two-stage MOT algorithms in Chapter 3. Nevertheless, we still noted the metrics down in Table 3.2, so as to compare the performance with other algorithms and with the metrics that *GMTracker* achieved on other popular MOT datasets.

## Performance of object detection

In the experiment of object detection, we averaged the performances of implementing each object detector over *RGB-1/2/3*. Table A11 lists the specific model performance on each validation set. The YOLOX-x model performs similarly, with AP in the range 51 to 54 and AP<sub>50</sub> in 86 to 89. The Faster R-CNN model has an average precision between 53 to 59 (AP<sub>50</sub> between 85 to 92). The performances on *RGB-2* are relatively worse than the others, but the differences are limited. We also list the benchmark Average Precision (AP) of the model on the COCO dataset, as claimed by the model developers. The results validate that the models perform similarly on the three validation sets, so it is reasonable to use the averaged AP for discussion in Chapter 3. The comparison in Chapter 3 demonstrates that both models perform at a comparable level with the corresponding benchmarks that are stated by the model developers [87, 278]. Hence, in Chapter 3, we argue that the difficulty level of object detection on *GSD* is not significantly higher than other MOT datasets.

Table A11: Performances of object detectors on *RGB-1/2/3* of *GSD* respectively. We use AP, AP<sub>50</sub> and AP<sub>75</sub> as the metrics. The first column gives the model of the object detector, and the second column indicates the validation set. The last row of each detector section, written as “average”, is the averaged performance of the models on the three validation sets, calculated without a weight. Performance of object detectors on *GSD*, evaluated by AP, AP<sub>50</sub> and AP<sub>75</sub>. All the values are averaged over the metrics of the three models that tested on the camera *RGB-1/2/3* respectively.

Detector	Validation Set	AP	AP <sub>50</sub>	AP <sub>75</sub>
YOLOX-x				
	RGB-1	53.7	88.3	57.7
	RGB-2	51.0	86.0	54.7
	RGB-3	62.4	87.5	71.7
	Average	55.7	87.3	61.4
Faster R-CNN				
	RGB-1	58.2	91.5	65.9
	RGB-2	53.3	87.9	58.0
	RGB-3	56.0	85.8	65.7
	Average	55.8	88.4	63.2
YOLOX-x	COCO[87]	59.2	86.3	61.9
Faster R-CNN	COCO[278]	40.2	60.9	43.8

MOT Metrics correlation with data characteristics

In Chapter 3, we highlight two primary challenges presented by this dataset: irregular movements and significant appearance changes exhibited by a majority of the objects. To further investigate the impact of dataset characteristics, we conducted MOT performance evaluation over fixed-duration periods. These periods were determined using a rolling window of 14 days, with a stride of 7 days. We measured changes in object appearances by calculating the  $\chi^2$  distance of each color spectrum at 2 p.m. daily, so as to limit the daily illumination variance. Differences in object locations were quantified by averaging the location changes of the same objects, based on bounding box (bbox) coordinates. Additionally, we computed the average lengths of trajectories (TL) within each period, which is defined as the number of bbox annotations. To assess the correlation between these indicators and the MOT metrics derived from DeepSORT, which incorporates both location and appearance features during data association, we analyzed the results.

The correlation values, depicted in figure A9, demonstrate the extent of influence on the metrics. It is evident that color changes exert a significant impact on all performance metrics. Notably, the False Positive (FP) rate of detected-and-associated bounding boxes is particularly affected, indicating an increased number of missed object matches when relying on appearance-based association across frames. The influence of object movement variations, while comparatively less pronounced than color changes, is more strongly correlated with tracklet identification performance. This observation aligns with the challenges posed by sudden position changes resulting from horticulture activity interruptions and the sparsity of data collection. The correlations between FP rate and redness, as well as between FP rate and trajectory length (TL), reach values of 0.7, underscoring their significant contributions to overall performance. Moreover, longer trajectories exhibit a strong correlation with reduced Multiple Object Tracking Accuracy (MOTA) and precision, indi-

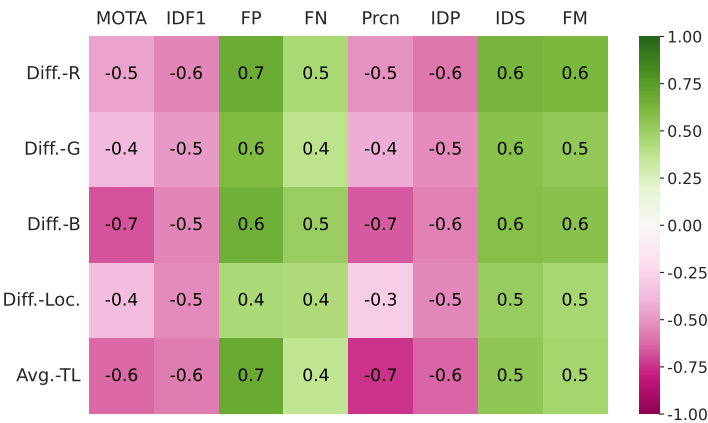


Figure A9: Correlation analysis among the performance metrics of DeepSORT and trajectories characteristics. The values in the grid indicate the exact value of correlation between the MOT metric from the horizontal axis, which is mentioned on the top of the figure, and the characteristics indicator from the vertical axis and listed on the left. The color of each grid is defined by the correlations, according to the scale shown on the right.

cating a substantial negative impact resulting from the extended duration of the tracking task. Taken together, these correlations provide evidence of the challenges introduced by the *GSD*: i) the appearance change of objects over the long period and ii) the irregular movements recorded in the sparse frames.

## Comparisons of trajectories

figure A10 gives a more abstracted comparison of trajectories in *GSD-2021-RGB-1* and *MOT20-01*, using the ground-truth annotations of the first and last observation of each track. The visualization demonstrates that a majority of objects in both sequences experienced changes in their locations, with object movements in *MOT20-01* generally being more substantial. This observation, when compared with Figure 3.4, further supports the distinctive pattern of movement exhibited by objects in the *GSD*: predominantly static yet with sudden and irregular changes.

The Euclidean distances and their ratios to the area of the initial observation (1<sup>st</sup> bbox) in both sequences exhibit similar distributions, indicating that the scales of movements in *GSD-2021-RGB-1* and *MOT20-01* are comparable.

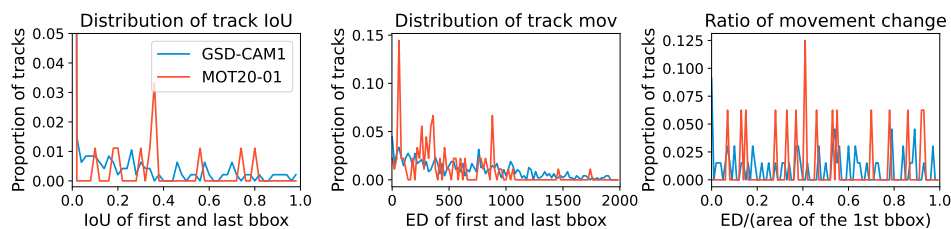


Figure A10: Quantitative comparisons of tracks in *GSD-2021-RGB-1* and *MOT20-01*. The 1<sup>st</sup> plot shows the IoU distribution of the first and last bbox of each track. The 2<sup>nd</sup> plot illustrates the IoU Euclidean distance (ED) of the first and last bbox of each track. The 3<sup>rd</sup> plot presents the proportion of the Euclidean distance (ED) of the first and last bbox of each track to the size of the first bbox.

## A.7 Further experimental results of Chapter 4

### Feature of env-only experiment

As a preliminary analysis, we computed the correlations of the environmental records and the average Brix of each harvest. We considered the environmental records from the week before the harvest. The results in Figure A11 indicate a strong correlation between temperatures (from the surface, leaves, plants, outdoor, etc.), radiations, watering, and cyclic lighting with the Brix expectation.

We grouped the attributes according to their absolute correlation with the Brix, using thresholds of 0.3 and 0.5. Although the data under each category were highly correlated, the data selection did not negatively affect the prediction performances. Neither did it significantly improve accuracy. Hence, we included only the best-performing model in the result section.

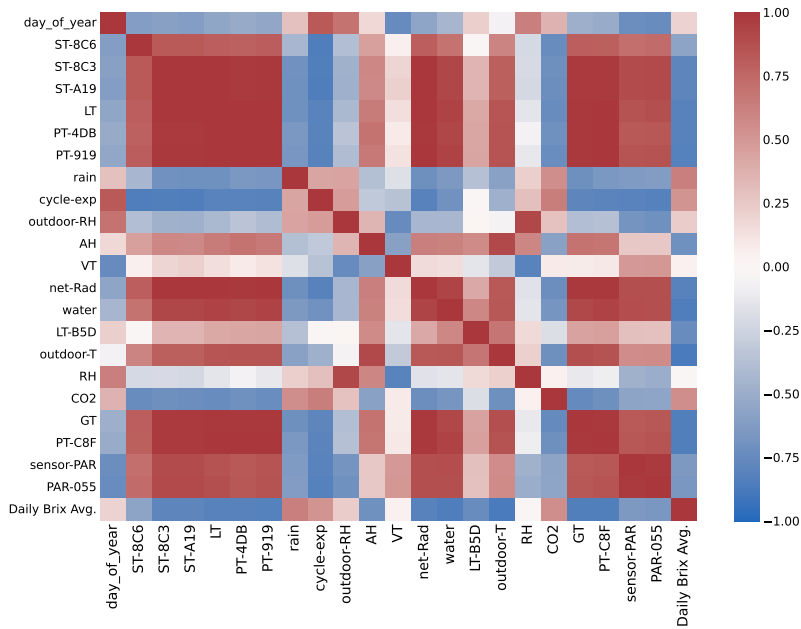


Figure A11: Correlations of each combination of features and the features with the daily averaged Brix. The color indicates the definition of each abbreviated name on the x-axis, which can be referred to Table 2.1. The heatmap gives the level of correlation.

## Regression example of env-only experiment

With the aggregated Brix information that was only predicted with environment records, we used the leave-one-out method to split the train and validation data. Figure A12 illustrates the prediction results from two selected models.

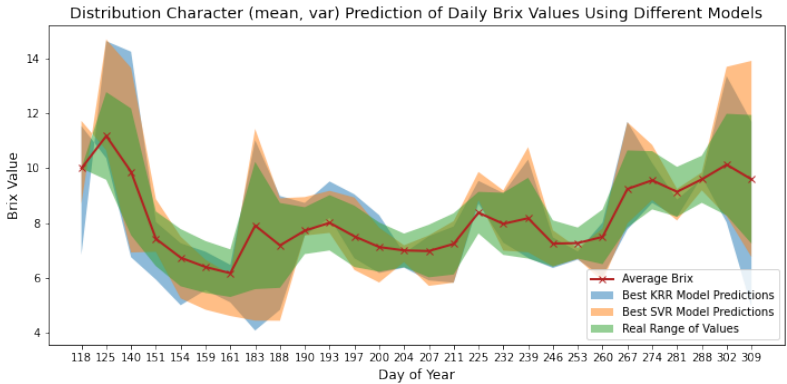


Figure A12: Brix distribution prediction over the measurement period. The distribution is illustrated by the (predicted) average and standard deviation of the Brix at the harvest day. The best model is selected by the minimum average RMSE over all data points, under various groups of feature selection.

## Validation on Further Datasets

In 2022 and 2024, we undertook two additional rounds of growth monitoring and Brix measurements in two greenhouses, continuing our efforts to gather comprehensive data on plant development and fruit quality. These new datasets allowed us to examine growth trends and sugar content dynamics over multiple years, providing a robust foundation for further analysis. Leveraging these observations, we conducted a detailed validation study aimed at corroborating the findings and methodologies presented in Chapter 4. This extended investigation not only reinforces the initial results but also offers new insights into the consistency and applicability of the proposed approaches under varying greenhouse conditions.

### Data exploration

In addition to displaying measurements that are used to label strawberry image segments, we examine the distribution of all the relevant measurements over the years. With an increased sample size, we observe in Figure A13 a generally smaller standard deviation, though with a slight shift in values. These statistics underscore the importance of larger sample sizes to ensure that the label distributions to be relatively more representative.

Table A12: Availability of datasets

Year	Mean	Std	Image Data	Climate Data
2021	8.16	1.73	Y	Y
2022	7.25	1.57	Y	N
2024: treatment 1	7.27	1.24	Y	Y
2024: treatment 2	7.33	1.20	Y	Y

### Overall performance comparison

The validation process utilizes data collected in 2024 and is carried out using two distinct approaches. First, we employed the best-performing or most representative models identified in Chapter 4. Second, we re-trained models using datasets from 2021 and 2022 to evaluate performance under alternative training conditions. These validation results, combined with the benchmarks detailed in Chapter 4, are summarized in Table A13, providing a comprehensive comparison of model performance.

As shown in the table, simply applying the model to the new dataset can result in an RMSE worse than the standard deviation of the testing set, indicating that the model performs worse than a constant guess. The model performs particularly poorly when it only considers environmental features, as it becomes less robust due to the limited amount of data. However, incorporating a ground truth of the distribution significantly reduces the prediction error, as the anchor points are updated to reflect the new dataset. In conclusion, when using models on new data, it is always recommended to sample a few data points from the new set. This can either contribute to retraining the model or provide additional insights into the distribution.

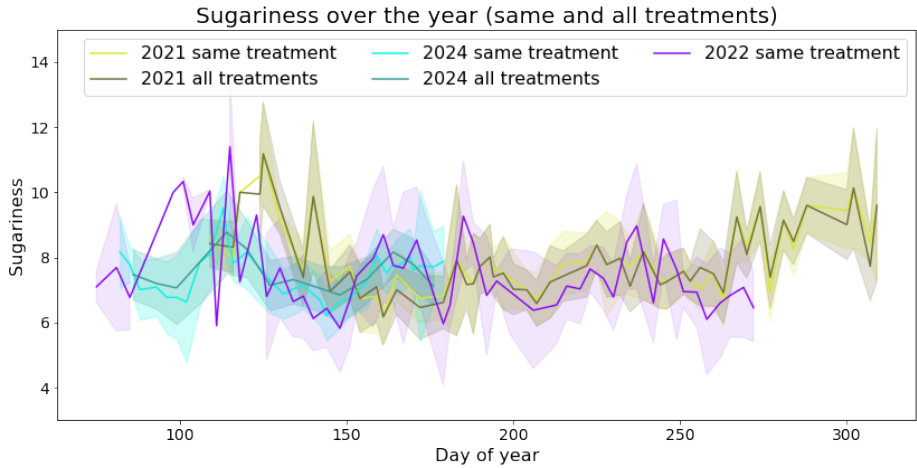


Figure A13: Statistics of the aggregated Brix measurements over the datasets. The x-axis depicts the dates using the "day of year". The y-axis shows the Sugariness measured in °Brix. The solid lines represent the mean values and the contour indicates the standard deviation of the samples of the day.

Table A13: performance comparison all image-brix datasets

	2021 (Paper)	model in paper, validated on '24	re-train by '21+'22, validated on '24
Std. of testing set	1.73	1.15	1.15
Image-only model	1.33	1.24	1.13
Env.-only model	1.27* (new: 1.22)	2.35	1.47
Image+mean(gt)	1.13	1.18	1.04
Image+quantiles(gt)	1.13	1.10	1.00
Image+mean(pred)	1.15*	1.61	1.35
Image+quantiles(pred)	1.10*	2.03	1.65

## A.8 Further experimental results of Chapter 6

### Loss in the training and validation set

Below, Figure A14 presents the average regret in the training set and validation set at the last epoch of training. By comparing the values from each pair of plots for the same model architecture, we can observe that deeper architectures tend not to overfit the training set but rather encounter challenges such as getting trapped in local optima during training.

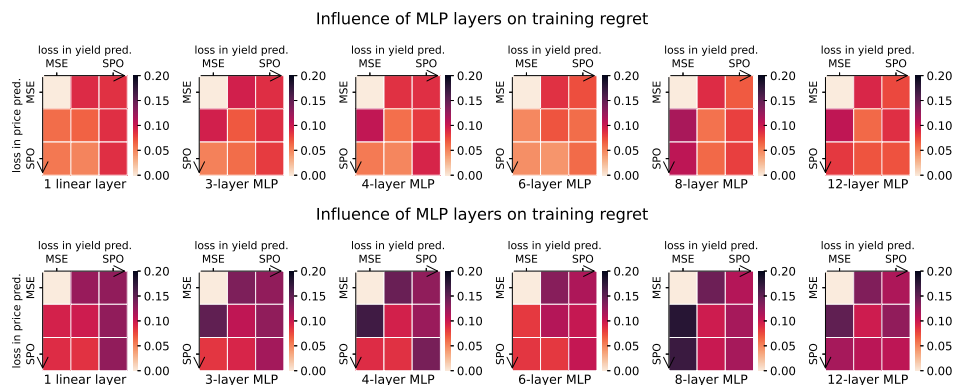


Figure A14: Regret of models varies with MLP layer numbers and loss function combinations. The X-axis label means the loss for yield prediction training and the Y-axis label is that for price. The value in the middle boxes ("M+S") used loss =  $0.5 \cdot \text{MSE} + 0.5 \cdot \text{SPO}$ . The upper left corners value 0 because no regret was calculated when the models were both trained by MSE. All plots share the same scale of color illustration but are different than Figures 6.1 and 6.4. *Upper plot*: results are collected from the last epoch's training performance over five seeds. *Lower plot*: results of the ending validation performance.

## More weighted aggregations of training losses

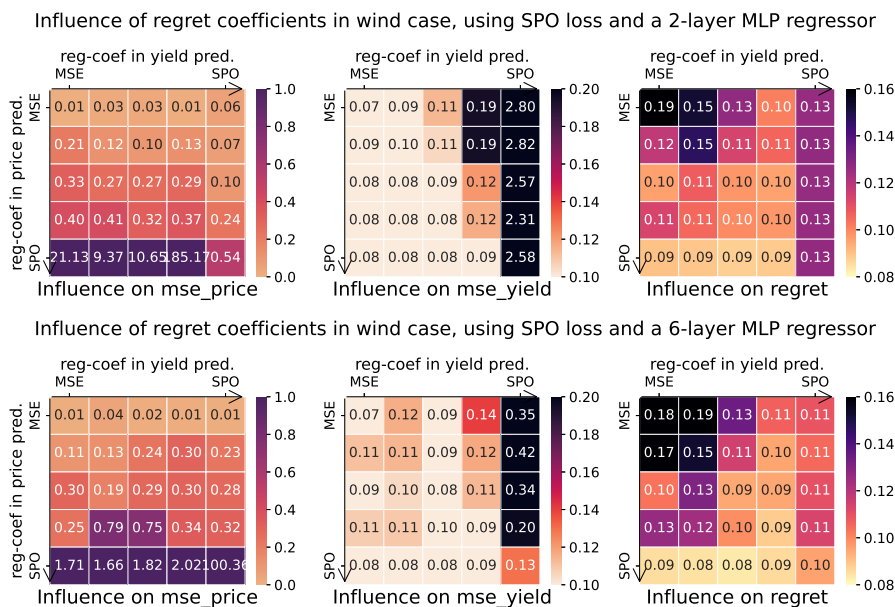


Figure A15: Prediction accuracies and regret under different combinations of MSE and SPO loss of 6-layer MLP models. The intermediate blocks are results from  $0.9 \cdot \text{MSE} + 0.1 \cdot \text{SPO}$ ,  $0.5 \cdot 0.5$ , and  $0.1 \cdot 0.9$  (in the same direction as the arrows). Values were averaged over five testing sets.

As presented by the detailed color maps in Figure A15, the MSE of price prediction boosts in both cases when the MSE was not involved and regularized the training. Such bias in prediction results in a higher incentive to invest (more) in the market with unknown parameters, which usually results in better regret according to the third plot of each row. However, the 2-layer model failed to give an accurate prediction about the yield, whereas the MSE on yield is similar to its prediction-focused learning (PFL) benchmarks in more cases in the 6-layer model. Our empirical findings lead us to conclude that the decline in decision-making performance primarily stemmed from inaccuracies in predicting the yield. The yield serves as both a pivotal decision parameter and a constraint within the problem. The results indicate that a reliable estimation of the feasible region is crucial to implement SPO loss properly.

The gradients from different training purposes show similar scales, as illustrated in Figure A16. Upon examining the boxplot, it becomes apparent that the gradients share similar scales. Consequently, weight combinations such as 0.1-0.9, 0.5-0.5, and 0.9-0.1 are all viable options for aggregating the losses.

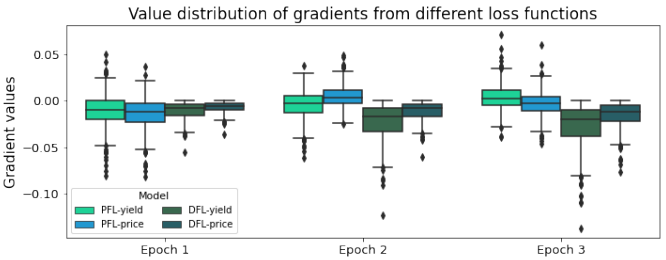


Figure A16: Gradient from prediction-focused learning (PFL) and decision-focused learning (DFL) in the first three epochs. A 3-layer MLP model is used as an example.

## Further layer comparison of models

In Chapter 6, we mainly compared the weight and bias in the MLP layers. In fact, the hyperparameters of the (fully-connected) RNN layer and self-attention module are also very similar when the model were trained under MSE or SPO loss. The comparisons of full models are depicted in Figure A17.

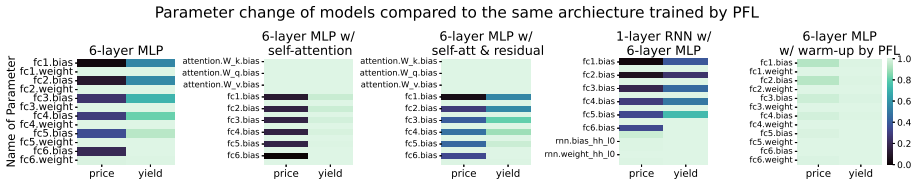


Figure A17: Similarity of parameters between models with 6 MLP layers ("fc") when trained on SPO loss and MSE. The last plot compares DFL models with pre-trained on PFL with the same PFL model. We used the cosine similarity of corresponding layer parameters, noted on the Y-axis. The color scale ranges from 0 to 1, as shown on the shared color bar on the right.



### Influence of model architecture on regret

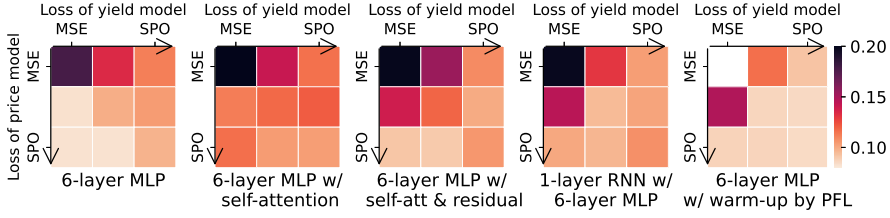


Figure A18: Regret of models varies with MLP layer numbers and loss function combinations. The models are aligned with the above layer parameter comparison plot. The X-axis label means the loss for yield prediction training and the Y-axis label is that for price. All plots share the same color scale (shown on the right). Results are collected over five seeds.

## Specific performance of models after pre-training

In Chapter 6, we discussed the final regret of models that were pre-trained on PFL and compared with the performance with original benchmarks and with single-model pre-training, as shown in the upper plot of Figure A19. In fact, by applying the pre-training strategy, the prediction accuracies were also better maintained. The lower plot of Figure A19 presents the MSE of price and yield predictions respectively. This ensured that the decision was made under more realistic constraints. This is one of the reasons why the regret is further improved by the pre-training strategy.

### Average regret on test data with different warm-up strategies

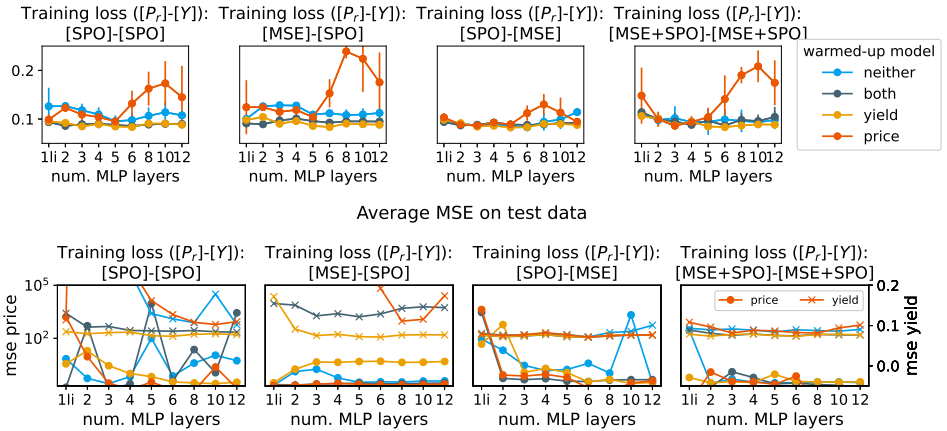


Figure A19: Comparison of the effect of the transfer learning strategy by the regret (upper) and MSE (lower), after switching to different combinations of training losses. The color bars show which model(s) was/were pre-training on the MSE. The x-axis shows the depth of the MLP model. In the lower figure, the values on the left y-axis indicate the MSE of price prediction and the right y-axis is for the MSE of yield. The line color is shared with the upper plot, and the marker identifies the metric. Results collected over five seeds.

## Enhanced yield prediction is a strategic jump start

Since all the above results were achieved from models trained together from scratch, we looked into the change of accuracy and regret during training. We observed from Figure A20 that SPO loss converged more smoothly after the MSE on yield prediction mostly converged. Hence a faster convergence of yield prediction is preferred. This can be achieved by using models with higher capacity, e.g. with a deeper architecture or using self-attention to highlight the features, which enhance the learning efficiency for smaller/shallower models. Moreover, the changes during the training process also inspire us to pre-train the model or use accelerating techniques such as adding self-attention.

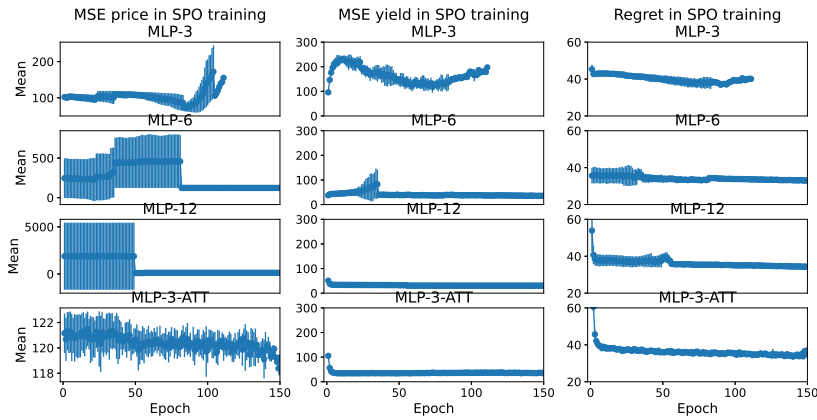


Figure A20: The training process of models, represented by MSE and SPO loss. Models are trained simultaneously unless early-stopped. The x-axis is the shared epoch number. The first two lines are from MLP models with 3 and 8 layers respectively. The last line is a 3-layer MLP model with a self-attention module after the input layer. All training used SPO loss solely for both price and yield prediction models. Results collected over five seeds.

Interestingly, the training logs also reveal a key area for improvement: models tend to influence each other when using the joint SPO loss, potentially leading them down incorrect or sub-optimal paths. To mitigate this risk, training should commence with weights that aren't entirely random but are initialized in a more informed and reasonable manner.

By checking the losses of PFL and DFL during the training progress of different models, we found that smoother loss convergence and better performance of DFL highly rely on the stable prediction performance of at least one model. According to Figure A19, models that were trained solely on SPO loss from scratch can easily perform biased predictions, which do not always benefit the final decision-making.

## Transfer learning and self-attention can accelerate DFL

The computational time is also a crucial criterion of DFL due to the extensive duration required for performing differentiation within the objective function. Table A14 compares the computational time of DFL by the above-mentioned regressors. The findings suggest that the training time of MLP models can be significantly reduced by integrating atten-

tion modules or by pre-training based on MSE, of which the converging time until early stopping is on average  $< 0.01$  hour. A similar effect was also been demonstrated by [169] in their setting, that the pre-training on (an easier) part of the main task facilitates the overall training efficiency.

Table A14: Average training time of models when they were both trained on MSE or SPO. The time is averaged over models of the same initial layer as presented in Figure A19, plus the warmed-up MLP models. "pre\_on\_[x]" indicates a pre-training process was applied on model [x]. The total time is counted until both models converge on the loss or reach the epoch limit.

Regressor	DFL Time [h]
mlp	1.23
rnn+mlp	1.37
atti+mlp	1.38
atti+res+mlp	1.87
mlp+pre_on_both	0.76
mlp+pre_on_price	1.34
mlp+pre_on_yield	0.49

## Influence of learning rate

As illustrated in Figure A21, the learning rate of model training has limited influences on the final decision-making performance. We used a learning rate of 0.002 in the experiments as a balanced choice, aiming to avoid excessively slow training while also mitigating the risk of getting stuck in local minima. According to the first subplot, it is sufficient in training adequate PFL models. Moreover, the consistent trend observed in the regret plot suggests that using this same learning rate for pre-training is likewise viable.

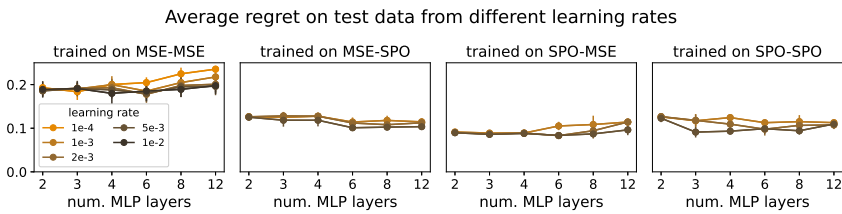


Figure A21: Influence of learning rate on the testing regret on various depths of MLP models. The y-axis gives the average regret on the test sets. Results collected over three seeds.

## Influence of optimizers

We also compared the performance of the optimizer on PFL and DFL progress. As illustrated in Figure A22, stochastic gradient descent (SGD) and Adam showed similar performance, with SGD demonstrating greater stability. Considering that SGD is more robust to hyperparameter changes, we used only SGD for the experiments in Chapter 6.

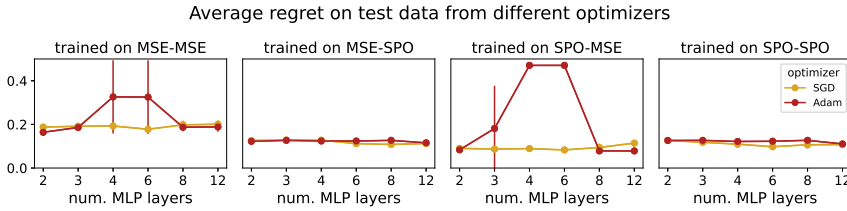


Figure A22: Influence of optimizer on the testing regret on various depths of MLP models. All models used a learning rate of 0.002. The y-axis gives the average regret on the test sets. Results collected over three seeds.

## Positioning of the Self-Attention Module

Figure A23 compares decision-making performance of models with different ways of integrating the self-attention layer. The results suggest that substituting the input layer with a self-attention layer is not optimal – while it speeds up progress, it fails to reach the global optimum. Utilizing residuals from the initial MLP layer or relying solely on the input data can prevent entrapment in local optima while retaining the acceleration advantages.

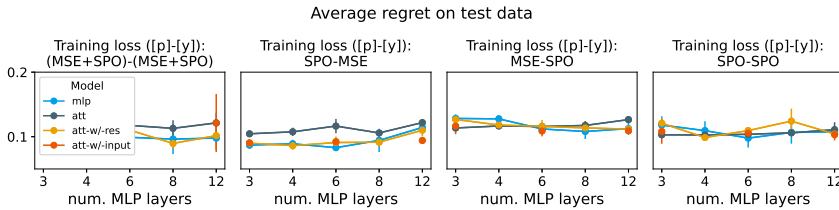


Figure A23: Influence of integrating self-attention layer on the testing regret for models with different MLP depths. The y-axis gives the average regret. In the legend, “w/-res” means the second MLP layer receives the outputs from the self-attention layer and the residuals from the first MLP layer, and “w/-input” means the second MLP layer receives the self-attention outputs and original input data. Results collected over five seeds.

## Influence of using a subset of data

As shown in Figure A24, the models perform similarly when it was trained on the full dataset, with also the discrepancy exists. Nonetheless, under a 4-hour limit in training

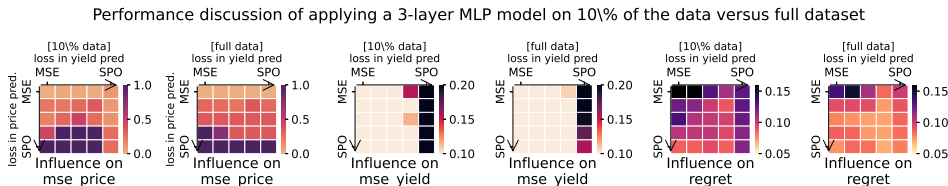


Figure A24: Comparison of MSE on price and yield prediction, and regret, when the same model was trained on 10% or on the full dataset. The intermediate blocks are results from  $0.9 \cdot \text{MSE} + 0.1 \cdot \text{SPO}$ ,  $0.5 \cdot 0.5$ , and  $0.1 \cdot 0.9$  (in the same direction as the arrows). Values were averaged over five testing sets.

time, less than 1/3 of the model can finish the DFL process. Particularly, when the models were both trained exclusively on SPO loss, none of the training processes ended within 4 hours. Hence, using a 10% is sufficient to represent the findings, but the experiments can be done more efficiently.

## Case-specific benchmarking

Figures A25 and A26 illustrate the performance of the selected models for the "wind-ori" and "tomato-c" cases. Prediction errors –  $(P_{r,pred} - P_{r,real})$  in white and  $(y_{pred} - y_{real})$  in yellow – are shown together with the resulting regrets, collected over multiple trials using 3-layer MLPs with different seeds on the full dataset. Four baselines were compared: (a) empirical decisions made from the latest parameters, (b) decisions from predictive models directly, (c) using standard PFL to train price/yield models on MSE, and (d) price/yield models trained on cost-weighted MSE. To quantify the involvement of SPO loss in training, we used the ratio of SPO and MSE loss, denoted as  $w_1 : w_2$  (as described in Section 6.4).

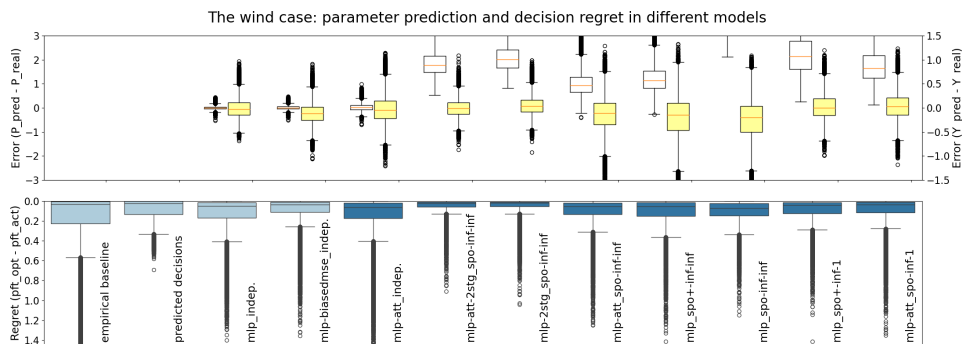


Figure A25: Prediction error of market price and yield and resulting regret in the "wind-ori" case. Upper boxes: price (white) and yield (yellow); lower boxes: baseline models (light blue) and experiments (dark blue). Model configurations (architecture, DFL loss: SPO/SPO+,  $w_1 : w_2$ , "2stg" = PFL warm-up) are next to boxplots of regrets.

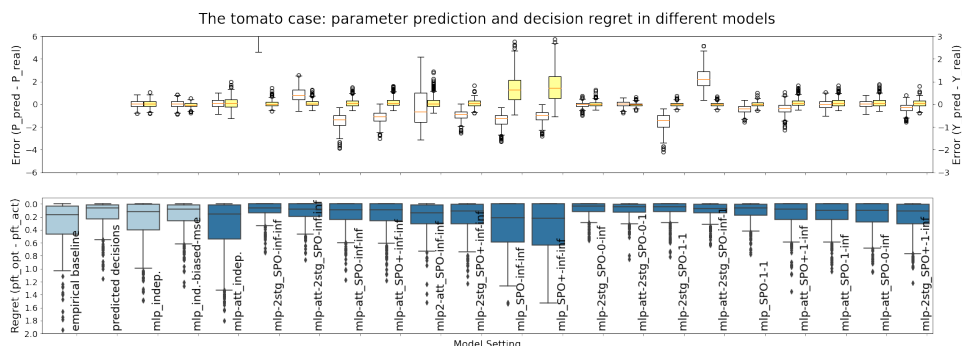


Figure A26: Prediction error of market price and yield and resulting regret in the "tomato-c" case, where the contract price is considered as always no lower than market price. Legends and notes are the same as above.

# Bibliography

- [1] Abasi, S.; Minaei, S.; Jamshidi, B.; and Fathi, D. 2018. Dedicated non-destructive devices for food quality measurement: A review. *Trends in Food Science & Technology*, 78(May): 197–205.
- [2] Abeytilakarathna, P.; Fonseka, R.; Eswara, J.; and Wijethunga, K. 2013. Relationship between total solid content and red, green and blue colour intensity of strawberry (*Fragaria x ananassa* Duch.) fruits. *Journal of Agricultural Sciences*, 8(2): 82.
- [3] Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- [4] Aday, M. S.; Temizkan, R.; Büyükcan, M. B.; and Caner, C. 2013. An innovative technique for extending shelf life of strawberry: Ultrasound. *LWT – Food Science and Technology*, 52(2): 93–101.
- [5] Aguiar, N.; Gupta, V.; and Khargonekar, P. P. 2020. A Real Options Market-Based Approach to Increase Penetration of Renewables. *IEEE Transactions on Smart Grid*, 11(2): 1691–1701.
- [6] Agulheiro-Santos, A. C.; Ricardo-Rodrigues, S.; Laranjo, M.; Melgão, C.; and Velázquez, R. 2022. Non-destructive prediction of total soluble solids in strawberry using near infrared spectroscopy. *Journal of the Science of Food and Agriculture*, 102(11): 4866–4872.
- [7] Ahumada, O.; and Villalobos, J. R. 2009. Application of planning models in the agri-food supply chain: A review. *European Journal of Operational Research*, 196(1): 1–20.
- [8] Akay, B.; Karaboga, D.; and Akay, R. 2022. A comprehensive survey on optimizing deep learning models by metaheuristics. *Artificial Intelligence Review*, 55(2): 829–894.
- [9] Amodio, M. L.; Chaudhry, M. M. A.; and Colelli, G. 2019. Spectral and Hyperspectral Technologies as an Additional Tool to Increase Information on Quality and Origin of Horticultural Crops. *Agronomy*, 10(1): 7.
- [10] Amoriello, T.; Ciccoritti, R.; and Ferrante, P. 2022. Prediction of Strawberries' Quality Parameters Using Artificial Neural Networks. *Agronomy*, 12(4).
- [11] An, Q.; Wang, K.; Li, Z.; Song, C.; Tang, X.; and Song, J. 2022. Real-time monitoring method of strawberry fruit growth state based on YOLO improved model. *IEEE Access*, 10: 124363–124372.

- [12] Andriluka, M.; Roth, S.; and Schiele, B. 2008. People-tracking-by-detection and people-detection-by-tracking. In *2008 IEEE Conference on computer vision and pattern recognition*, 1–8.
- [13] Anjum, S.; and Gurari, D. 2020. CTMC: Cell Tracking with Mitosis Detection Dataset Challenge. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, volume 2020-June, 4228–4237.
- [14] Ariza-Sentís, M.; Vélez, S.; Martínez-Peña, R.; Baja, H.; and Valente, J. 2024. Object detection and tracking in Precision Farming: a systematic review. *Computers and Electronics in Agriculture*, 219: 108757.
- [15] Avsar, E.; Bulus, K.; Saridas, M. A.; and Kapur, B. 2018. Development of a cloud-based automatic irrigation system: A case study on strawberry cultivation. In *2018 7th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, 1–4.
- [16] Ayala-Silva, T.; Winterstein, R. J.; Meerow, A. W.; Winterstein, M.; Cervantes, C.; and Brown, J. S. 2005. Determination of color and fruit traits of half-sib families of mango (*Mangifera indica* L.). In *Proceedings of the Florida State Horticultural Society*, volume 118, 253–257.
- [17] Ayoub Shaikh, T.; Rasool, T.; and Rasheed Lone, F. 2022. Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. *Computers and Electronics in Agriculture*, 198: 107119.
- [18] Azodanlou, R.; Darbellay, C.; Luisier, J.-L.; Villettaz, J.-C.; and Amadò, R. 2003. Quality Assessment of Strawberries (*Fragaria* Species). *Journal of Agricultural and Food Chemistry*, 51(3): 715–721.
- [19] Baghdasaryan, L.; Melikbekyan, R.; Dolmajain, A.; and Hobbs, J. 2022. Deep density estimation based on multi-spectral remote sensing data for in-field crop yield forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014–2023.
- [20] Bali, N.; and Singla, A. 2021. Deep Learning Based Wheat Crop Yield Prediction Model in Punjab Region of North India. *Applied Artificial Intelligence*, 35(15): 1304–1328.
- [21] Bao, R.; Al-Shakarji, N. M.; Bunyak, F.; and Palaniappan, K. 2021. Dmnet: Dual-stream marker guided deep network for dense cell segmentation and lineage tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3361–3370.
- [22] Bargoti, S.; and Underwood, J. P. 2017. Image Segmentation for Fruit Detection and Yield Estimation in Apple Orchards. *Journal of Field Robotics*, 34(6): 1039–1060.
- [23] Belda, S.; Pipia, L.; Morcillo-Pallarés, P.; and Verrelst, J. 2020. Optimizing Gaussian Process Regression for Image Time Series Gap-Filling and Crop Monitoring. *Agronomy*, 10(5): 618.

- [24] Bellemare, M. F.; and Bloem, J. R. 2018. Does contract farming improve welfare? A review. *World Development*, 112: 259–271.
- [25] Bergmann, P.; Meinhardt, T.; and Leal-Taixe, L. 2019. Tracking Without Bells and Whistles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 941–951.
- [26] Bernardin, K.; and Stiefelhausen, R. 2008. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP Journal on Image and Video Processing*, 2008: 1–10.
- [27] Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; and Upcroft, B. 2016. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, volume 2016-August, 3464–3468.
- [28] Bezryadin, S.; Bourov, P.; and Ilinih, D. 2007. Brightness Calculation in Digital Image Processing. *International Symposium on Technologies for Digital Photo Fulfillment*, 1(1): 10–15.
- [29] Bezryadin, S.; Bourov, P.; and Ilinih, D. 2007. Brightness Calculation in Digital Image Processing. *International Symposium on Technologies for Digital Photo Fulfillment*, 1(1): 10–15.
- [30] Bhargava, A.; and Bansal, A. 2021. Fruits and vegetables quality evaluation using computer vision: A review. *Journal of King Saud University-Computer and Information Sciences*, 33(3): 243–257.
- [31] Blakeney, M. 2019. *Food Loss and Food Waste*. Edward Elgar Publishing.
- [32] Block, C.; Neumann, D.; and Weinhardt, C. 2008. A market mechanism for energy allocation in micro-chp grids. In *Proc. of the 41st annual Hawaii international conference on system sciences (HICSS 2008)*, 172–172.
- [33] Borsa, N. 2023. Tomato Prices. <http://novataborsa.com/%D1%86%D0%B5%D0%BD%D0%B8-%D0%B4%D0%BE%D0%BC%D0%B0%D1%82%D0%B8/>. Accessed on: August 2023.
- [34] Bos-Brouwers, H. E. J.; Soethoudt, J. M.; Vollebregt, H.; and van der Burgh, M. 2015. Monitor voedselverspilling: update Monitor voedselverspilling 2009-2013 & mogelijkheden tot (zelf) monitoring van voedselverspilling door de keten heen. Technical Report 486902, Wageningen University & Research.
- [35] Brasó, G.; and Leal-Taixé, L. 2020. Learning a Neural Solver for Multiple Object Tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6246–6256.
- [36] Brealey, R. A. 2000. Stock prices, stock indexes and index funds. *Bank of England Quarterly Bulletin*.



- [37] Brecht, J.; Loaza, F.; Nunes, M.; Emond, J.; Uysal, I.; Badia, F.; Wells, J.; and Saenz, J. 2016. Reducing strawberry waste and losses in the postharvest supply chain via intelligent distribution management. *Acta Horticulturae*, 1120(1120): 253–260.
- [38] Brooks, C.; and Garrett, I. 2002. Can we explain the dynamics of the UK FTSE 100 stock and stock index futures markets? *Applied Financial Economics*, 12(1): 25–31.
- [39] Bunyak, F.; Palaniappan, K.; Nath, S. K.; Baskin, T.; and Dong, G. 2006. Quantitative cell motility for in vitro wound healing using level set-based active contour tracking. In *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006.*, 1040–1043.
- [40] Cacchiani, V.; Caprara, A.; and Toth, P. 2010. Solving a real-world train-unit assignment problem. *Mathematical programming*, 124: 207–231.
- [41] Cai, J.; Tang, Y.; Yan, K.; Harrison, A. P.; Xiao, J.; Lin, G.; and Lu, L. 2021. Deep Lesion Tracker: Monitoring Lesions in 4D Longitudinal Imaging Studies. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15154–15164.
- [42] Cai, W.; Liang, Y.; Liu, X.; Feng, J.; and Wu, Y. 2024. Msgnet: Learning multi-scale inter-series correlations for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11141–11149.
- [43] Caixeta-Filho, J. V.; van Swaay-Neto, J. M.; and Wagemaker, A. d. P. 2002. Optimization of the production planning and trade of lily flowers at Jan de Wit Company. *Interfaces*, 32(1): 35–46.
- [44] Calleja, E.; Ilbery, B.; and Mills, P. 2012. Agricultural change and the rise of the British strawberry industry, 1920–2009. *Journal of Rural Studies*, 28(4): 603–611.
- [45] Cao, J.; Pang, J.; Weng, X.; Khirodkar, R.; and Kitani, K. 2023. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9686–9696.
- [46] Catalá, L. P.; Moreno, M. S.; Blanco, A. M.; and Bandoni, J. A. 2016. A bi-objective optimization model for tactical planning in the pome fruit industry supply chain. *Computers and Electronics in Agriculture*, 130: 128–141.
- [47] Centre for the Promotion of Imports from developing countries (CBI). 2021. Entering the European market for fresh strawberries (Market Entry). <https://www.cbi.eu/market-information/fresh-fruit-vegetables/strawberries/market-entry>. Last updated: 29 September 2021; Accessed: May 2025.
- [48] Centre for the Promotion of Imports from developing countries (CBI). 2021. The European market potential for strawberries. <https://www.cbi.eu/market-information/fresh-fruit-vegetables/strawberries/market-potential>. Last updated: 29 September 2021; Accessed: May 2025.

- [49] Chandler, C. K.; Folta, K.; Dale, A.; Whitaker, V. M.; and Herrington, M. 2012. Strawberry. In *Fruit Breeding*, 305–325. Boston, MA: Springer US.
- [50] Chang, X.; Wu, J.; Sun, H.; and Yan, X. 2023. A Smart Predict-then-Optimize method for dynamic green bike relocation in the free-floating system. *Transportation Research Part C: Emerging Technologies*, 153: 104220.
- [51] Chen, F.; Tang, Y.-N.; and Shen, M.-Y. 2011. Coordination control of greenhouse environmental factors. *International Journal of Automation and Computing*, 8(2): 147–153.
- [52] Chen, L.; Ai, H.; Shang, C.; Zhuang, Z.; and Bai, B. 2017. Online multi-object tracking with convolutional neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, volume 2017-September, 645–649.
- [53] Chen, W.; Zhang, H.; Mehlawat, M. K.; and Jia, L. 2021. Mean-variance portfolio optimization using machine learning-based stock price prediction. *Applied Soft Computing*, 100: 106943.
- [54] Chen, Y.; Lee, W. S.; Gan, H.; Peres, N.; Fraisse, C.; Zhang, Y.; and He, Y. 2019. Strawberry Yield Prediction Based on a Deep Neural Network Using High-Resolution Aerial Orthoimages. *Remote Sensing*, 11(13): 1584.
- [55] Cho, W.; Na, M.; Kim, S.; and Jeon, W. 2019. Automatic prediction of brix and acidity in stages of ripeness of strawberries using image processing techniques. In *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, 1–4.
- [56] Choi, H. G.; Moon, B. Y.; and Kang, N. J. 2015. Effects of LED light on the production of strawberry during cultivation in a plastic greenhouse and in a growth chamber. *Scientia Horticulturae*, 189: 22–31.
- [57] Clark, E.; Qiao, Z.; and Wong, W.-K. 2016. Theories of risk: Testing investor behavior on the taiwan stock and stock index futures markets. *Economic Inquiry*, 54(2): 907–924.
- [58] Conejo, A. J.; Carrión, M.; Morales, J. M.; et al. 2010. *Decision making under uncertainty in electricity markets*, volume 1. Springer.
- [59] Connor, J. T.; Martin, R. D.; and Atlas, L. E. 1994. Recurrent neural networks and robust time series prediction. *IEEE transactions on neural networks*, 5(2): 240–254.
- [60] Corallo, A.; Latino, M. E.; and Menegoli, M. 2018. From Industry 4.0 to Agriculture 4.0: A Framework to Manage Product Data in Agri-Food Supply Chain for Voluntary Traceability, A framework proposed. *International Journal of Nutrition and Food Engineering*, 11(5): 126–130.
- [61] Correia, P.; Pestana, M.; Martinez, F.; Ribeiro, E.; Gama, F.; Saavedra, T.; and Palencia, P. 2011. Relationships between strawberry fruit quality attributes and crop load. *Scientia Horticulturae*, 130(2): 398–403.

- [62] Cuturi, M.; and Blondel, M. 2017. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, 894–903.
- [63] Darby-Dowman, K.; Barker, S.; Audsley, E.; and Parsons, D. 2000. A two-stage stochastic programming with recourse model for determining robust planting plans in horticulture. *Journal of the Operational Research Society*, 51(1): 83–89.
- [64] Das, J.; Cross, G.; Qu, C.; Makineni, A.; Tokekar, P.; Mulgaonkar, Y.; and Kumar, V. 2015. Devices, systems, and methods for automated monitoring enabling precision agriculture. In *2015 IEEE international conference on automation science and engineering (CASE)*, 462–469.
- [65] Dave, A.; Khurana, T.; Tokmakov, P.; Schmid, C.; and Ramanan, D. 2020. TAO: A Large-Scale Benchmark for Tracking Any Object. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12350 LNCS, 436–454. Springer Science and Business Media Deutschland GmbH.
- [66] Delac, K.; Grgic, M.; and Grgic, S. 2005. Independent comparative study of PCA, ICA, and LDA on the FERET data set. *International Journal of Imaging Systems and Technology*, 15(5): 252–260.
- [67] Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; and Leal-Taixé, L. 2020. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*.
- [68] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255.
- [69] Diamond, S.; and Boyd, S. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83): 1–5.
- [70] Díaz-Galián, M. V.; Torres, M.; Sanchez-Pagán, J. D.; Navarro, P. J.; Weiss, J.; and Egea-Cortines, M. 2021. Enhancement of strawberry production and fruit quality by blue and red LED lights in research and commercial greenhouses. *South African Journal of Botany*, 140: 269–275.
- [71] Dileep, B.; Grover, R.; and Rai, K. 2002. Contract farming in tomato: An economic analysis. *Indian Journal of Agricultural Economics*, 57(2): 197–210.
- [72] Dong, J.; Gruda, N.; Lam, S. K.; Li, X.; and Duan, Z. 2018. Effects of elevated CO<sub>2</sub> on nutritional quality of vegetables: a review. *Frontiers in plant science*, 9: 924.
- [73] Donti, P.; Amos, B.; and Kolter, J. Z. 2017. Task-based end-to-end model learning in stochastic optimization. *Advances in neural information processing systems*, 30.
- [74] Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; and Meng, H. 2023. StrongSORT: Make deepsort great again. *IEEE Transactions on Multimedia*.

- [75] Ekanayake, B.; Wong, J. K.-W.; Fini, A. A. F.; and Smith, P. 2021. Computer vision-based interior construction progress monitoring: A literature review and future research directions. *Automation in Construction*, 127: 103705.
- [76] Elhariri, E.; El-Bendary, N.; and Mahmoud Saleh, S. 2023. Strawberry-DS: Dataset of annotated strawberry fruits images with various developmental stages. *Data in Brief*, 48: 109165.
- [77] Elik, A.; Yanik, D. K.; Guzelsoy, N. A.; Yavuz, A.; and Gogus, F. 2019. Strategies to Reduce Post-Harvest Losses for Fruits and Vegetables. *International Journal of Scientific and Technological Research*.
- [78] Elmachoub, A. N.; and Grigas, P. 2022. Smart “predict, then optimize”. *Management Science*, 68(1): 9–26.
- [79] Fabbri, M.; Braso, G.; Maugeri, G.; Cetintas, O.; Gasparini, R.; Osep, A.; Calderara, S.; Leal-Taixe, L.; and Cucchiara, R. 2021. MOTSynth: How Can Synthetic Data Help Pedestrian Detection and Tracking? In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10829–10839.
- [80] Fang, S.-L.; Kuo, Y.-H.; Kang, L.; Chen, C.-C.; Hsieh, C.-Y.; Yao, M.-H.; and Kuo, B.-J. 2022. Using sigmoid growth models to simulate greenhouse tomato growth and development. *Horticulturae*, 8(11): 1021.
- [81] Fathollahi, F.; Ponnambalam, K.; Huang, Y.; and Karray, F. 2020. Agent-Based Modeling to Simulate Real-World Prices: A Strawberry Market Study. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 3606–3611.
- [82] Food; and of the United Nations (FAO), A. O. 2015. Global initiative on food loss and waste. Technical Report i7657e, FAO, Rome, Italy. Accessed: May 2023.
- [83] Fujinaga, T.; Yasukawa, S.; and Ishii, K. 2020. Tomato Growth State Map for the Automation of Monitoring and Harvesting. *Journal of Robotics and Mechatronics*, 32(6): 1279–1291.
- [84] Gabaldon, J.; Zhang, D.; Lauderdale, L.; Miller, L.; Johnson-Roberson, M.; Barton, K.; and Shorter, K. A. 2022. Computer-vision object tracking for monitoring bottlenose dolphin habitat use and kinematics. *PloS one*, 17(2): e0254323.
- [85] Gao, J.; Li, P.; Chen, Z.; and Zhang, J. 2020. A Survey on Deep Learning for Multi-modal Data Fusion. *Neural Computation*, 32(5): 829–864.
- [86] Gao, Z.; Shao, Y.; Xuan, G.; Wang, Y.; Liu, Y.; and Han, X. 2020. Real-time hyperspectral imaging for the in-field estimation of strawberry ripeness with deep learning. *Artificial Intelligence in Agriculture*, 4: 31–38.
- [87] Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. YOLOX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

- [88] Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3354–3361.
- [89] Geladi, P.; Isaksson, H.; Lindqvist, L.; Wold, S.; and Esbensen, K. 1989. Principal component analysis of multivariate images. *Chemometrics and Intelligent Laboratory Systems*, 5(3): 209–220.
- [90] Gerhard, D.; and Moltchanova, E. 2022. A Richards growth model to predict fruit weight. *Australian & New Zealand Journal of Statistics*, 64(4): 413–421.
- [91] Gibbs, J. A.; Burgess, A. J.; Pound, M. P.; Pridmore, T. P.; and Murchie, E. H. 2019. Recovering wind-induced plant motion in dense field environments via deep learning and multiple object tracking. *Plant physiology*, 181(1): 28–42.
- [92] Gijssenij, A.; and Gevers, T. 2011. Color Constancy Using Natural Image Statistics and Scene Semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4): 687–698.
- [93] Given, N. K.; Venis, M. A.; and Gierson, D. 1988. Hormonal regulation of ripening in the strawberry, a non-climacteric fruit. *Planta*, 174(3): 402–406.
- [94] Goatman, K. A.; Whitwam, A. D.; Manivannan, A.; Olson, J. A.; and Sharp, P. F. 2003. Colour normalisation of retinal images. In *Proceedings of medical image understanding and analysis*, 49–52.
- [95] Goharian, M.; Bruwer, M.-J.; Jegatheesan, A.; Moran, G. R.; and MacGregor, J. F. 2007. A novel approach for EIT regularization via spatial and spectral principal component analysis. *Physiological Measurement*, 28(9): 1001–1016.
- [96] Gómez, A. H.; He, Y.; and Pereira, A. G. 2006. Non-destructive measurement of acidity, soluble solids and firmness of Satsuma mandarin using Vis/NIR-spectroscopy techniques. *Journal of Food Engineering*, 77(2): 313–319.
- [97] Gómez-Lagos, J. E.; González-Araya, M. C.; Soto-Silva, W. E.; and Rivera-Moraga, M. M. 2021. Optimizing tactical harvest planning for multiple fruit orchards using a metaheuristic modeling approach. *European journal of operational research*, 290(1): 297–312.
- [98] Gong, L.; Yu, M.; Jiang, S.; Cutsuridis, V.; and Pearson, S. 2021. Deep learning based prediction on greenhouse crop yield combined TCN and RNN. *Sensors*, 21(13): 4537.
- [99] Guo, S.; Wang, S.; Yang, Z.; Wang, L.; Zhang, H.; Guo, P.; Gao, Y.; and Guo, J. 2022. A Review of Deep Learning-Based Visual Multi-Object Tracking Algorithms for Autonomous Driving. *Applied Sciences*, 12(21): 10741.
- [100] Gustavsson, J.; Cederberg, C.; Sonesson, U.; van Otterdijk, R.; and Meybeck, A. 2011. Global food losses and food waste: Extent, causes and prevention. Technical report, Food and Agriculture Organization of the United Nations, Rome, Italy. Accessed: May 2025.

- [101] Gutiérrez, S.; Wendel, A.; and Underwood, J. 2019. Ground based hyperspectral imaging for extensive mango yield estimation. *Computers and Electronics in Agriculture*, 157: 126–135.
- [102] Gutiérrez, S.; Wendel, A.; and Underwood, J. 2019. Spectral filter design based on in-field hyperspectral imaging and machine learning for mango ripeness estimation. *Computers and Electronics in Agriculture*, 164: 104890.
- [103] Hamuda, E.; Mc Ginley, B.; Glavin, M.; and Jones, E. 2018. Improved image processing-based crop detection using Kalman filtering and the Hungarian algorithm. *Computers and electronics in agriculture*, 148: 37–44.
- [104] Harel, B.; Edan, Y.; and Perlman, Y. 2022. Optimization Model for Selective Harvest Planning Performed by Humans and Robots. *Applied Sciences*, 12(5).
- [105] Hayashi, S.; Yamamoto, S.; Saito, S.; Ochiai, Y.; Nagasaki, Y.; and Kohno, Y. 2013. Structural Environment Suited to the Operation of a Strawberry-harvesting Robot Mounted on a Travelling Platform. *Engineering in Agriculture, Environment and Food*, 6(1): 34–40.
- [106] He, J.; Huang, Z.; Wang, N.; and Zhang, Z. 2021. Learnable Graph Matching: Incorporating Graph Partitioning with Deep Feature Learning for Multiple Object Tracking. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5295–5305.
- [107] He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.
- [108] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*.
- [109] Hersbach, H.; Bell, B.; Berrisford, P.; Biavati, G.; Horányi, A.; Muñoz Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Rozum, I.; Schepers, D.; Simmons, A.; Soci, C.; Dee, D.; and Thépaut, J.-N. 2023. ERA5 hourly data on single levels from 1940 to present. Accessed on: July 2024.
- [110] Hidaka, K.; Dan, K.; Miyoshi, Y.; Imamura, H.; Takayama, T.; Kitano, M.; Sameshima, K.; and Okimura, M. 2016. Twofold increase in strawberry productivity by integration of environmental control and movable beds in a large-scale greenhouse. *Environmental Control in Biology*, 54(2): 79–92.
- [111] Hopf, A.; Boote, K. J.; Oh, J.; Guan, Z.; Agehara, S.; Shelia, V.; Whitaker, V. M.; Asseng, S.; Zhao, X.; and Hoogenboom, G. 2022. Development and improvement of the CROPGRO-Strawberry model. *Scientia Horticulturae*, 291: 110538.
- [112] Hu, X.; Lee, J. C.; and Lee, J. H. 2023. Predict+Optimize for packing and covering LPs with unknown parameters in constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3987–3995.

- [113] Hu, X.; Lee, J. C.; and Lee, J. H. 2023. Two-Stage Predict+Optimize for MILPs with Unknown Parameters in Constraints. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- [114] Huang, S.; Tang, L.; Hupy, J. P.; Wang, Y.; and Shao, G. 2021. A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing. *Journal of Forestry Research*, 32(1): 1–6.
- [115] Huang, Y. H.; and Te Lin, T. 2019. High-throughput image analysis framework for fruit detection, localization and measurement from video streams. In *2019 ASABE Annual International Meeting*, 1.
- [116] Hueth, B.; Ligon, E.; Wolf, S.; and Wu, S. 1999. Incentive Instruments in Fruit and Vegetable Contracts: Input Control, Monitoring, Measuring, and Price Risk. *Applied Economic Perspectives and Policy*, 21(2): 374–389.
- [117] Hui, A.; Y Soo, Y. O.; T Mok, V. C.; Sing Lawrence Wong, K.; and Hui Patrick Kwan TW Leung Y Soo Vincent CT Mok Lawrence KS Wong, A. C. 2007. Diagnostic value and safety of long-term video-EEG monitoring. *Hong Kong Medical Journal*, 13(3).
- [118] Jannatizadeh, A.; Rezaei, M.; Rohani, A.; Lawson, S.; and Fatahi, R. 2023. Towards modeling growth of apricot fruit: finding a proper growth model. *Horticulture, Environment, and Biotechnology*, 64(2): 209–222.
- [119] Janurianti, N. M. D.; Utama, I. M. S.; and Gunam, I. B. W. 2021. Colour and quality of strawberry fruit (*Fragaria x ananassa* Duch.) at different levels of maturity. *SEAS (Sustainable Environment Agricultural Science)*, 5(1): 22–28.
- [120] Jha, S.; Chopra, S.; and Kingsly, A. 2007. Modeling of color values for nondestructive evaluation of maturity of mango. *Journal of Food Engineering*, 78(1): 22–26.
- [121] Jhana, N. 2019. Hourly energy demand generation and weather. <https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather>. Accessed on: October, 2023.
- [122] Jiménez-Bravo, D. M.; Murciego, Á. L.; Mendes, A. S.; San Blás, H. S.; and Bajo, J. 2022. Multi-Object Tracking in Traffic Environments: A Systematic Literature Review. *Neurocomputing*.
- [123] Jol, C.; Wen, J.; and Van Gemert, J. 2023. Non-Destructive Infield Quality Estimation of Strawberries using Deep Architectures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 515–524.
- [124] Kafkas, E.; Koşar, M.; Paydaş, S.; Kafkas, S.; and Başer, K. 2007. Quality characteristics of strawberry genotypes at different maturation stages. *Food chemistry*, 100(3): 1229–1236.



- [125] Kaiser, H. M.; and Aplan, J. 1989. DSSP: a model of production and marketing decisions on a Midwestern crop farm. *Applied Economic Perspectives and Policy*, 11(2): 157–169.
- [126] Kaiser, H. M.; Riha, S. J.; Wilks, D. S.; Rossiter, D. G.; and Sampath, R. 1993. A farm-level analysis of economic and agronomic impacts of gradual climate warming. *American journal of agricultural economics*, 75(2): 387–398.
- [127] Kaldmäe, H.; Kikas, A.; Arus, L.; and Libek, A.-V. 2013. Genotype and microclimate conditions influence ripening pattern and quality of blackcurrant (*Ribes nigrum* L.) fruit. *Zemdirbyste-Agriculture*, 100(2): 167–174. Accessed: May 2025.
- [128] kalman\_pub. 2022. Benchmark results of ByteTrack on MOT17. Last submitted on November 23, 2021.
- [129] Karpathy, A. 2022. NanoGPT. <https://github.com/karpathy/nanoGPT>.
- [130] Kelly, K.; Madden, R.; Emond, J. P.; and do Nascimento Nunes, M. C. 2019. A novel approach to determine the impact level of each step along the supply chain on strawberry quality. *Postharvest Biology and Technology*, 147: 78–88.
- [131] Keogh, E.; and Ratanamahatana, C. A. 2005. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7: 358–386.
- [132] Kim, I. S.; Choi, H. S.; Yi, K. M.; Choi, J. Y.; and Kong, S. G. 2010. Intelligent visual surveillance – a survey. *International Journal of Control, Automation and Systems*, 8(5): 926–939.
- [133] Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.
- [134] Kirschen, D. S.; and Strbac, G. 2018. *Fundamentals of power system economics*. John Wiley & Sons.
- [135] Klinbumrung, N.; and Teerachaichayut, S. 2018. Quantification of acidity and total soluble solids in guavas by near infrared hyperspectral imaging. In *AIP Conference Proceedings*, volume 2030, 020209.
- [136] Kotary, J.; Fioretto, F.; Van Hentenryck, P.; and Wilder, B. 2021. End-to-end constrained optimization learning: A survey. *arXiv:2103.16378*.
- [137] Krishnamurthy, D.; Uckun, C.; Zhou, Z.; Thimmapuram, P. R.; and Botterud, A. 2018. Energy Storage Arbitrage Under Day-Ahead and Real-Time Price Uncertainty. *IEEE Transactions on Power Systems*, 33(1): 84–93.
- [138] Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.



- [139] Kumar, S. V.; and Bhasker, R. 2023. An Artificial Rabbit Optimized Neural Network-Based Hybrid DC-AC Microgrid Energy Management System. In *2023 7th International Conference On Computing, Communication, Control And Automation (IC-CUBE)*, 1–7.
- [140] Lamb, N.; and Chuah, M. C. 2018. A Strawberry Detection System Using Convolutional Neural Networks. In *2018 IEEE International Conference on Big Data (Big Data)*, 2515–2520.
- [141] Lambert, G. F.; Lasserre, A. A. A.; Ackerman, M. M.; Sánchez, C. G. M.; Rivera, B. O. I.; and Azzaro-Pantel, C. 2014. An expert system for predicting orchard yield and fruit quality and its impact on the Persian lime supply chain. *Engineering Applications of Artificial Intelligence*, 33: 21–30.
- [142] Lavanya, M.; and Santhi, B. 2021. Hungarian optimization technique based efficient resource allocation using clustering unbalanced estimated cost matrix. *Journal of Ambient Intelligence and Humanized Computing*, 12: 5525–5540.
- [143] Le Centre Technique Interprofessionnel des Fruits et Légumes. 2017. Code couleur fraise. <https://www.ctifl.fr/code-couleur-fraise>. Last access on May 2025.
- [144] Lee, D.-J.; Archibald, J. K.; and Xiong, G. 2011. Rapid Color Grading for Fruit Quality Evaluation Using Direct Color Mapping. *IEEE Transactions on Automation Science and Engineering*, 8(2): 292–302.
- [145] Lee, J.; Guo, J.; Choi, J. K.; and Zukerman, M. 2015. Distributed Energy Trading in Microgrids: A Game-Theoretic Model and Its Equilibrium Analysis. *IEEE Transactions on Industrial Electronics*, 62(6): 3524–3533.
- [146] Lee, M. A.; Monteiro, A.; Barclay, A.; Marcar, J.; Miteva-Neagu, M.; and Parker, J. 2020. A framework for predicting soft-fruit yields and phenology using embedded, networked microsensors, coupled weather models and machine-learning techniques. *Computers and Electronics in Agriculture*, 168: 105103.
- [147] Lezoche, M.; Hernandez, J. E.; Alemany Díaz, M. d. M. E.; Panetto, H.; and Kacprzyk, J. 2020. Agri-food 4.0: A survey of the supply chains and technologies for the future agriculture. *Computers in Industry*, 117: 103187.
- [148] Li, T.; Li, Y.; and Qian, Y. 2016. Improved Hungarian algorithm for assignment problems of serial-parallel systems. *Journal of systems engineering and electronics*, 27(4): 858–870.
- [149] Li, X.; Wang, K.; Wang, W.; and Li, Y. 2010. A multiple object tracking method using Kalman filter. In *The 2010 IEEE international conference on information and automation*, 1862–1866.
- [150] Li, Z.; and Snavely, N. 2018. Learning Intrinsic Image Decomposition from Watching the World. In *Computer Vision and Pattern Recognition (CVPR)*.

- [151] Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.
- [152] Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.
- [153] Liu, C.; Liu, W.; Lu, X.; Ma, F.; Chen, W.; Yang, J.; and Zheng, L. 2014. Application of Multispectral Imaging to Determine Quality Attributes and Ripeness Stage in Strawberry Fruit. *PLoS ONE*, 9(2): e87818.
- [154] Liu, F.; Snetkov, L.; and Lima, D. 2017. Summary on fruit identification methods: A literature review. In *2017 3rd International Conference on Economics, Social Science, Arts, Education and Management Engineering (ESSAEME 2017)*.
- [155] Liu, G.; Xu, Y.; and Tomsovic, K. 2016. Bidding Strategy for Microgrid in Day-Ahead Market Based on Hybrid Stochastic/Robust Optimization. *IEEE Transactions on Smart Grid*, 7(1): 227–237.
- [156] Liu, H.; and Grigas, P. 2021. Risk bounds and calibration for a smart predict-then-optimize method. *Advances in Neural Information Processing Systems*, 34: 22083–22094.
- [157] Liu, H.; and Grigas, P. 2022. Online Contextual Decision-Making with a Smart Predict-then-Optimize Method. *arXiv preprint*.
- [158] Liu, Q.; Wei, K.; Xiao, H.; Tu, S.; Sun, K.; Sun, Y.; Pan, L.; and Tu, K. 2019. Near-Infrared Hyperspectral Imaging Rapidly Detects the Decay of Postharvest Strawberry Based on Water-Soluble Sugar Analysis. *Food Analytical Methods*, 12(4): 936–946.
- [159] Loesdau, M.; Chabrier, S.; and Gabillon, A. 2014. Hue and Saturation in the RGB Color Space. In Elmoataz, A.; Lezoray, O.; Nouboud, F.; and Mammass, D., eds., *Image and Signal Processing*, 203–212.
- [160] Lowe, T. J.; and Preckel, P. V. 2004. Decision Technologies for Agribusiness Problems: A Brief Review of Selected Literature and a Call for Research. *Manufacturing & Service Operations Management*, 6(3): 201–208.
- [161] Luiten, J.; Osep, A.; Dendorfer, P.; Torr, P.; Geiger, A.; Leal-Taixé, L.; and Leibe, B. 2021. HOTA: A Higher Order Metric for Evaluating Multi-object Tracking. *International Journal of Computer Vision*, 129(2): 548–578.
- [162] Luo, C.; Yang, X.; and Yuille, A. 2021. Exploring simple 3d multi-object tracking for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10488–10497.
- [163] Luo, W.; Xing, J.; Milan, A.; Zhang, X.; Liu, W.; and Kim, T.-K. 2021. Multiple object tracking: A literature review. *Artificial Intelligence*, 293: 103448.

- [164] Luo, W.; Zhao, X.; and Kim, T.-K. 2014. Multiple object tracking: A review. *arXiv preprint arXiv:1409.7618*, 1: 1.
- [165] Mahendra, O.; Pardede, H. F.; Sustika, R.; and Suryo Kusumo, R. B. 2018. Comparison of Features for Strawberry Grading Classification with Novel Dataset. In *2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 7–12.
- [166] Mancini, M.; Mazzoni, L.; Gagliardi, F.; Balducci, F.; Duca, D.; Toscano, G.; Mezzetti, B.; and Capocasa, F. 2020. Application of the Non-Destructive NIR Technique for the Evaluation of Strawberry Fruits Quality Parameters. *Foods*, 9(4): 441.
- [167] Mandi, J.; Bucarey, V.; Tchomba, M. M. K.; and Guns, T. 2022. Decision-Focused Learning: Through the Lens of Learning to Rank. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 14935–14947.
- [168] Mandi, J.; Kotary, J.; Berden, S.; Mulamba, M.; Bucarey, V.; Guns, T.; and Fioretto, F. 2023. Decision-Focused Learning: Foundations, State of the Art, Benchmark and Future Opportunities. *arXiv preprint*.
- [169] Mandi, J.; Stuckey, P. J.; Guns, T.; et al. 2020. Smart predict-and-optimize for hard combinatorial optimization problems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 1603–1610.
- [170] Mangawati, A.; Leesan, M.; Aradhya, H. R.; et al. 2018. Object Tracking Algorithms for video surveillance applications. In *2018 International Conference on Communication and Signal Processing (ICCSP)*, 0667–0671.
- [171] Marques, E. J. N.; de Freitas, S. T.; Pimentel, M. F.; and Pasquini, C. 2016. Rapid and non-destructive determination of quality parameters in the ‘Tommy Atkins’ mango using a novel handheld near infrared spectrometer. *Food chemistry*, 197: 1207–1214.
- [172] Mavridou, E.; Vrochidou, E.; Papakostas, G. A.; Pachidis, T.; and Kaburlasos, V. G. 2019. Machine vision systems in precision agriculture for crop farming. *Journal of Imaging*, 5(12): 89.
- [173] Meijering, E.; Dzyubachyk, O.; Smal, I.; and van Cappellen, W. A. 2009. Tracking in cell and developmental biology. *Seminars in Cell & Developmental Biology*, 20(8): 894–902.
- [174] Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; and Schindler, K. 2016. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*.
- [175] Mirhaji, H.; Soleymani, M.; Asakereh, A.; and Mehdizadeh, S. A. 2021. Fruit detection and load estimation of an orange orchard using the YOLO models through simple approaches in different imaging and illumination conditions. *Computers and Electronics in Agriculture*, 191: 106533.

- [176] Mishra, P.; Sytsma, M.; Chauhan, A.; Polder, G.; and Pekkeriet, E. 2022. All-in-one: a spectral imaging laboratory system for standardised automated image acquisition and real-time spectral model deployment. *Analytica Chimica Acta*, 1190: 339235.
- [177] Mizrach, A. 2008. Ultrasonic technology for quality evaluation of fresh fruit and vegetables in pre- and postharvest processes. *Postharvest Biology and Technology*, 48(3): 315–330.
- [178] Montero, T. M.; Mollá, E. M.; Esteban, R. M.; and López-Andréu, F. J. 1996. Quality attributes of strawberry during ripening. *Scientia Horticulturae*, 65(4): 239–250.
- [179] MOTChallenge. 2023. MOTChallenge – Multiple Object Tracking Benchmark. Accessed on August 22, 2023.
- [180] Mouafik, M.; Fouad, M.; and El Aboudi, A. 2024. Machine Learning Methods for Predicting *Argania spinosa* Crop Yield and Leaf Area Index: A Combined Drought Index Approach from Multisource Remote Sensing Data. *AgriEngineering*, 6(3): 2283.
- [181] Muangprathub, J.; Boonnam, N.; Kajornkasirat, S.; Lekbangpong, N.; Wanichsombat, A.; and Nillaor, P. 2019. IoT and agriculture data analysis for smart farm. *Computers and Electronics in Agriculture*, 156: 467–474.
- [182] Mula, J.; Peidro, D.; Díaz-Madroñero, M.; and Vicens, E. 2010. Mathematical programming models for supply chain production and transport planning. *European Journal of Operational Research*, 204(3): 377–390.
- [183] Munera, S.; Amigo, J. M.; Blasco, J.; Cubero, S.; Talens, P.; and Aleixos, N. 2017. Ripeness monitoring of two cultivars of nectarine using VIS-NIR hyperspectral reflectance imaging. *Journal of Food Engineering*, 214: 29–39.
- [184] Munhoz, J. R.; and Morabito, R. 2014. Optimization approaches to support decision making in the production planning of a citrus company: A Brazilian case study. *Computers and Electronics in Agriculture*, 107: 45–57.
- [185] Najafizadeh, L.; and Froehlich, J. E. 2018. A Feasibility Study of Using Google Street View and Computer Vision to Track the Evolution of Urban Accessibility. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, 340–342.
- [186] Nandi, C. S.; Tudu, B.; and Koley, C. 2016. A Machine Vision Technique for Grading of Harvested Mangoes Based on Maturity and Quality. *IEEE Sensors Journal*, 16(16): 6387–6396.
- [187] Nardello, I. C.; Cantillano, R. F. F.; Ribeiro, J. A.; Vignolo, G. K.; Malgarim, M. B.; and Mello-Farias, P. C. d. 2021. Quality of strawberries subjected to pre-harvest UV-C radiation. *Pesquisa Agropecuária Brasileira*, 56: e01590.
- [188] Nate Gruver, S. Q., Marc Finzi; and Wilson, A. G. 2023. Large Language Models Are Zero Shot Time Series Forecasters. In *Advances in Neural Information Processing Systems*.

- [189] Ncube, D. 2020. The importance of contract farming to small-scale farmers in Africa and the implications for policy: A review scenario. *The Open Agriculture Journal*, 14(1).
- [190] Nuske, S.; Wilshusen, K.; Achar, S.; Yoder, L.; Narasimhan, S.; and Singh, S. 2014. Automated Visual Yield Estimation in Vineyards. *Journal of Field Robotics*, 31(5): 837–860.
- [191] OCSORTpublic. 2022. Benchmark results of Observation-Centric SORT with public detections on MOT17. Last submitted on March 29, 2022.
- [192] Osman, Y.; Dennis, R.; and Elgazzar, K. 2021. Yield estimation and visualization solution for precision agriculture. *Sensors*, 21(19): 6657.
- [193] Osterloff, J.; Nilssen, I.; and Nattkemper, T. W. 2016. A computer vision approach for monitoring the spatial and temporal shrimp distribution at the LoVe observatory. *Methods in Oceanography*, 15-16: 114–128.
- [194] Otsuka, K.; Nakano, Y.; and Takahashi, K. 2016. Contract farming in developed and developing countries. *Annual Review of Resource Economics*, 8(1): 353–376.
- [195] Papageorgiou, E.; Aggelopoulou, K.; Gemtos, T.; and Nanos, G. 2018. Development and Evaluation of a Fuzzy Inference System and a Neuro-Fuzzy Inference System for Grading Apple Quality. *Applied Artificial Intelligence*, 32(3): 253–280.
- [196] Pastell, M.; Madis, L.; and Bloch, V. 2022. Strawberry dataset for object detection.
- [197] Pattnaik, G.; Shrivastava, V. K.; and Parvathi, K. 2020. Transfer Learning-Based Framework for Classification of Pest in Tomato Plants. *Applied Artificial Intelligence*, 34(13): 981–993.
- [198] Payne, A.; Walsh, K.; Subedi, P.; and Jarvis, D. 2014. Estimating mango crop yield using image analysis using fruit at ‘stone hardening’ stage and night time imaging. *Computers and Electronics in Agriculture*, 100: 160–167.
- [199] Pérez-Borrero, I.; Marín-Santos, D.; Gegúndez-Arias, M. E.; and Cortés-Ancos, E. 2020. A fast and accurate deep learning method for strawberry instance segmentation. *Computers and Electronics in Agriculture*, 178: 105736.
- [200] Polder, G.; van der Heijden, G. W.; and Young, I. T. 2002. Spectral image analysis for measuring ripeness of tomatoes. *Transactions of the ASAE*, 45(4): 1155.
- [201] Porat, R.; Lichter, A.; Terry, L. A.; Harker, R.; and Buzby, J. 2018. Postharvest losses of fruit and vegetables during retail and in consumers’ homes: Quantifications, causes, and means of prevention. *Postharvest Biology and Technology*, 139: 135–149.
- [202] Rahimiyan, M.; and Baringo, L. 2016. Strategic Bidding for a Virtual Power Plant in the Day-Ahead and Real-Time Markets: A Price-Taker Robust Optimization Approach. *IEEE Transactions on Power Systems*, 31(4): 2676–2687.

- [203] Rahman, M. M.; Moniruzzaman, M.; Ahmad, M. R.; Sarker, B.; and Khurshid Alam, M. 2016. Maturity stages affect the postharvest quality and shelf-life of fruits of strawberry genotypes growing in subtropical regions. *Journal of the Saudi Society of Agricultural Sciences*, 15(1): 28–37.
- [204] Ramana, K. V. R.; Govindarajan, V. S.; Ranganna, S.; and Kefford, J. F. 1981. Citrus fruits – varieties, chemistry, technology, and quality evaluation. Part I: Varieties, production, handling, and storage. *C R C Critical Reviews in Food Science and Nutrition*, 15(4): 353–431.
- [205] Rees, D.; Farrell, G.; and Orchard, J. 2012. *Crop Post-Harvest: Science and Technology*. Wiley. ISBN 9780632057252.
- [206] Ren, L.; Mo, T.; Cheng, X.; and Li, X. 2023. Temporal–Frequency Attention Focusing for Time Series Extrinsic Regression via Auxiliary Task. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.
- [207] Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137–1149.
- [208] Research, W. U. . 1991. Jaarverslag 1991 / Proefstation voor de Fruitteelt. Technical Report 251587, Wageningen University & Research.
- [209] Rezaei, M.; and Liu, B. 2017. Food loss and waste in the food supply chain.
- [210] Richards, F. J. 1959. A flexible growth function for empirical use. *Journal of experimental Botany*, 10(2): 290–301.
- [211] Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance Measures and a Data Set for Multi-target, Multi-camera Tracking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9914 LNCS, 17–35. Springer Verlag.
- [212] Roshan, S. H.; Kazemitabar, J.; and Kheradmandian, G. 2022. Artificial Intelligence Aided Agricultural Sensors for Plant Frostbite Protection. *Applied Artificial Intelligence*, 36(1): 2031814.
- [213] Roshan Zamir, A.; Dehghan, A.; and Shah, M. 2012. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part II* 12, 343–356.
- [214] S., M. G. P.; and R., B. 2019. Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms. *Applied Artificial Intelligence*, 33(7): 621–642.
- [215] Sa, I.; Ge, Z.; Dayoub, F.; Upcroft, B.; Perez, T.; and McCool, C. 2016. DeepFruits: A Fruit Detection System Using Deep Neural Networks. *Sensors*, 16(8): 1222.

- [216] Safdarian, A.; Fotuhi-Firuzabad, M.; and Lehtonen, M. 2015. A medium-term decision model for disCos: Forward contracting and TOU pricing. *IEEE Transactions on Power Systems*, 30(3): 1143–1154.
- [217] Sakoe, H.; and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1): 43–49.
- [218] Sala, F.; Popescu, C. A.; Herbei, M. V.; and Rujescu, C. 2020. Model of color parameters variation and correction in relation to “Time-View” image acquisition effects in wheat crop. *Sustainability*, 12(6): 2470.
- [219] Samykanno, K.; Pang, E.; and Marriott, P. J. 2013. Genotypic and environmental effects on flavor attributes of ‘Albion’ and ‘Juliette’ strawberry fruits. *Scientia Horticulturae*, 164: 633–642.
- [220] Sang, L.; Xu, Y.; Long, H.; Hu, Q.; and Sun, H. 2022. Electricity Price Prediction for Energy Storage System Arbitrage: A Decision-Focused Approach. *IEEE Transactions on Smart Grid*, 13(4): 2822–2832.
- [221] Schanda, J. 2007. CIE colorimetry. *Colorimetry: Understanding the CIE system*, 25–78.
- [222] Schieffer, J.; and Vassalos, M. 2015. Risk and the use of contracts by vegetable growers. *Choices*, 30(3): 1–4.
- [223] Schofield, C.; Marchant, J.; White, R.; Brandl, N.; and Wilson, M. 1999. Monitoring Pig Growth using a Prototype Imaging System. *Journal of Agricultural Engineering Research*, 72(3): 205–210.
- [224] Sen, S.; Yu, L.; and Genc, T. 2006. A stochastic programming approach to power portfolio optimization. *Operations Research*, 54(1): 55–72.
- [225] Shafizadeh-Moghadam, H. 2021. Fully component selection: An efficient combination of feature selection and principal component analysis to increase model performance. *Expert Systems with Applications*, 186: 115678.
- [226] Shaikh, T. A.; Mir, W. A.; Rasool, T.; and Sofi, S. 2022. Machine learning for smart agriculture and precision farming: towards making the fields talk. *Archives of Computational Methods in Engineering*, 29(7): 4557–4597.
- [227] Shao, Y.; Wang, Y.; Xuan, G.; Gao, Z.; Hu, Z.; Gao, C.; and Wang, K. 2021. Assessment of Strawberry Ripeness Using Hyperspectral Imaging. *Analytical Letters*, 54(10): 1547–1560.
- [228] Sharma, A.; Jain, A.; Gupta, P.; and Chowdary, V. 2020. Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, 9: 4843–4873.



- [229] Sherrah, J.; and Gong, S. 2000. Tracking Discontinuous Motion Using Bayesian Inference. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 1843, 150–166. Springer Verlag.
- [230] Shifaz, A.; Pelletier, C.; Petitjean, F.; and Webb, G. I. 2023. Elastic similarity and distance measures for multivariate time series. *Knowledge and Information Systems*, 65(6): 2665–2698.
- [231] Shin, J.; and Lee, J. H. 2016. Multi-time scale procurement planning considering multiple suppliers and uncertainty in supply and demand. *Computers & Chemical Engineering*, 91: 114–126.
- [232] Sim, H. S.; Kim, D. S.; Ahn, M. G.; Ahn, S. R.; and Kim, S. K. 2020. Prediction of strawberry growth and fruit yield based on environmental and growth data in a greenhouse for soil cultivation with applied autonomous facilities. *Horticultural Science and Technology*, 38(6): 840–849.
- [233] Simpson, D. 2018. The Economic Importance of Strawberry Crops. In Hytönen, T.; Graham, J.; and Harrison, R., eds., *The Genomes of Rosaceous Berries and Their Wild Relatives*, Compendium of Plant Genomes, 1–7. Cham: Springer International Publishing. ISBN 978-3-319-76019-3.
- [234] Singh, S. 2005. Role of the state in contract farming in Thailand: experience and lessons. *ASEAN Economic Bulletin*, 22(2): 217–228.
- [235] Soosay, C.; and Kannusam, R. 2018. Scope for Industry 4.0 in Agri-food Supply Chains. In *Conference Paper Scope for industry 4.0 in agri-food supply chain*, 1–22.
- [236] Soto-Silva, W. E.; Nadal-Roig, E.; González-Araya, M. C.; and Pla-Aragones, L. M. 2016. Operational research models applied to the fresh fruit supply chain. *European Journal of Operational Research*, 251(2): 345–355.
- [237] Stenmarck, A.; Jansen, C.; Quested, T.; and Moates, G. 2016. Estimates of European food waste levels. *FUSIONS project*.
- [238] Sun, M.; Zhang, D.; Liu, L.; and Wang, Z. 2017. How to predict the sugariness and hardness of melons: A near-infrared hyperspectral imaging method. *Food Chemistry*, 218: 413–421.
- [239] Sundararaman, R.; De Almeida Braga, C.; Marchand, E.; and Pettre, J. 2021. Tracking Pedestrian Heads in Dense Crowd. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3864–3874.
- [240] Tang, B.; and Khalil, E. B. 2022. PyEPO: A PyTorch-based end-to-end predict-then-optimize library for linear and integer programming. *arXiv:2206.14234*.
- [241] Tang, B.; and Khalil, E. B. 2023. Multi-task predict-then-optimize. In *International Conference on Learning and Intelligent Optimization*, 506–522.



- [242] Taofik, A.; Ismail, N.; Gerhana, Y.; Komarujaman, K.; and Ramdhani, M. A. 2018. Design of smart system to detect ripeness of tomato and chili with new approach in data acquisition. In *IOP Conference Series: Materials Science and Engineering*, volume 288, 012018.
- [243] Taşkın, T.; and Bilgen, B. 2021. Optimization Models for Harvest and Production Planning in Agri-Food Supply Chain: A Systematic Review. *Logistics*, 5(3).
- [244] Terry, L. A.; Mena, C.; Williams, A.; Jenney, N.; and Whitehead, P. 2011. Fruit and vegetable resource maps: Mapping fruit and vegetable waste through the retail and wholesale supply chain. Technical report, WRAP, RC008.
- [245] Tian, H.; Wang, T.; Liu, Y.; Qiao, X.; and Li, Y. 2020. Computer vision technology in agricultural automation – A review. *Information Processing in Agriculture*, 7(1): 1–19.
- [246] Tian, X.; Yan, R.; Liu, Y.; and Wang, S. 2023. A smart predict-then-optimize method for targeted and cost-effective maritime transportation. *Transportation Research Part B: Methodological*, 172: 32–52.
- [247] Van Etten, A.; Hogan, D.; Manso, J. M.; Shermeyer, J.; Weir, N.; and Lewis, R. 2021. The Multi-Temporal Urban Development SpaceNet Dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6394–6403.
- [248] van Klompenburg, T.; Kassahun, A.; and Catal, C. 2020. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177: 105709.
- [249] van Staden, H. E.; Deprez, L.; and Boute, R. N. 2022. A dynamic “predict, then optimize” preventive maintenance approach using operational intervention data. *European Journal of Operational Research*, 302(3): 1079–1096.
- [250] Vandendriessche, T.; Vermeir, S.; Mayayo Martinez, C.; Hendrickx, Y.; Lammertyn, J.; Nicolaï, B.; and Hertog, M. 2013. Effect of ripening and inter-cultivar differences on strawberry quality. *LWT – Food Science and Technology*, 52(2): 62–70.
- [251] Vardhan, A.; Prastawa, M.; Sadeghi, N.; Vachet, C.; Piven, J.; and Gerig, G. 2015. Joint Longitudinal Modeling of Brain Appearance in Multimodal MRI for the Characterization of Early Brain Developmental Processes. In *LNCS*, volume 8682, 49–63. Springer, Cham.
- [252] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- [253] Ventura, M.; de Jager, A.; de Putter, H.; and Roelofs, F. P. 1998. Non-destructive determination of soluble solids in apple fruit by near infrared spectroscopy (NIRS). *Postharvest Biology and Technology*, 14(1): 21–27.

- [254] Verdouw, C.; Beulens, A.; Trienekens, J.; and Wolfert, J. 2010. Process modelling in demand-driven supply chains: A reference model for the fruit industry. *Computers and Electronics in Agriculture*, 73(2): 174–187.
- [255] Verma, S.; Mate, A.; Wang, K.; Taneja, A.; and Tambe, M. 2022. Case study: Applying decision focused learning in the real world. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.
- [256] Verrelst, J.; Rivera, J. P.; Moreno, J.; and Camps-Valls, G. 2013. Gaussian processes uncertainty estimates in experimental Sentinel-2 LAI and leaf chlorophyll content retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 86: 157–167.
- [257] Vu, Q. H.; Cen, L.; Ruta, D.; and Liu, M. 2022. Key Factors to Consider when Predicting the Costs of Forwarding Contracts. In *Proc. of the 17th Conference on Computer Science and Intelligence Systems, FedCSIS 2022*, 447–450.
- [258] Wan, S.; and Goudos, S. 2020. Faster R-CNN for multi-class fruit detection using a robotic vision system. *Computer Networks*, 168: 107036.
- [259] Wang, K.; Wilder, B.; Perrault, A.; and Tambe, M. 2020. Automatically learning compact quality-aware surrogates for optimization problems. *Advances in Neural Information Processing Systems*, 33: 9586–9596.
- [260] Wang, L.; Ahmad, F.; Luo, G.-l.; Umar, M.; and Kirikkaleli, D. 2022. Portfolio optimization of financial commodities with energy futures. *Annals of Operations Research*, 313(1): 401–439.
- [261] Wang, Y.; Yan, G.; Meng, Q.; Yao, T.; Han, J.; and Zhang, B. 2022. DSE-YOLO: Detail semantics enhancement YOLO for multi-stage strawberry detection. *Computers and electronics in agriculture*, 198: 107057.
- [262] Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; and Wang, S. 2020. Towards real-time multi-object tracking. In *European conference on computer vision*, 107–122.
- [263] Weber, M.; Xie, J.; Collins, M.; Zhu, Y.; Voigtlaender, P.; Adam, H.; Green, B.; Geiger, A.; Leibe, B.; Cremers, D.; et al. 2021. Step: Segmenting and tracking every pixel. *arXiv preprint arXiv:2102.11859*.
- [264] Wen, J.; Abeel, T.; and de Weerd, M. 2023. “How sweet are your strawberries?” : Predicting sugariness using non-destructive and affordable hardware. *Frontiers in Plant Science*, 14: 1160645.
- [265] Wen, J.; Dascălu, A.; Drashkov, D.; Marinski, L.; Bhatti, M. A.; Herrebout, M.; and Weinmann, M. 2023. Demonstration Abstracts of A Mobile Application for Near Real-Time Strawberry Quality Prediction. In *the 35th Benelux Conference on Artificial Intelligence and Machine Learning (BNAIC/BeNeLearn 2023)*. Demonstrations.
- [266] Wen, J.; de Weerd, M.; Abeel, T.; Verschoor, C.; Schuddebeurs, L.; Walraven, K.; Jochems, S.; and Theelen, V. 2023. Data underlying the research of Quality prediction of strawberries with RGB image segments. <https://data.4tu.nl/datasets/25eac1bc-82c9-47fc-98e4-ea5d81804c52/2>.

- [267] Wen, J.; Verschoor, C.; Abeel, T.; and de Weerd, M. M. 2024. Dark Images of "The Growing Strawberries Dataset". <https://data.4tu.nl/my/datasets/4250ca74-962a-4da6-a2bd-a763f0cea2a5>.
- [268] Wen, J.; Verschoor, C.; Abeel, T.; and de Weerd, M. M. 2024. Extension of "The Growing Strawberries Dataset". <https://data.4tu.nl/my/datasets/73424098-2e29-4cc0-9fb2-86e379c8ced0>.
- [269] Wen, J.; Verschoor, C.; Abeel, T.; and de Weerd, M. M. 2024. The Growing Strawberries Dataset. <https://data.4tu.nl/datasets/e3b31ece-cc88-4638-be10-8ccdd4c5f2f7>.
- [270] Wen, J.; Verschoor, C.; Abeel, T.; and de Weerd, M. M. 2024. Growth monitoring images of strawberries. <https://data.4tu.nl/my/collections/f2ad72ce-3e5b-429e-ba14-78eed903ce03>.
- [271] Wen, J.; Verschoor, C.; Jochems, S.; Lisanne Schuddebeur sand Theelen, V.; Walraven, K.; den Bakker, B.; Scholten, J.; Abeel, T.; and de Weerd, M. 2024. Data underlying the research of strawberry quality prediction with infield data. <https://data.4tu.nl/datasets/96a1cb2d-9f16-470f-8c46-744d5985a140>.
- [272] Wen, J.; Verschoor, C. R.; Feng, C.; Epure, I.-M.; Abeel, T.; and de Weerd, M. 2024. The Growing Strawberries Dataset: Tracking Multiple Objects with Biological Development over an Extended Period. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 7089–7099.
- [273] Wen, L.; Lei, Z.; Chang, M.-C.; Qi, H.; and Lyu, S. 2017. Multi-Camera Multi-Target Tracking with Space-Time-View Hyper-graph. *International Journal of Computer Vision*, 122(2): 313–333.
- [274] Weng, S.; Yu, S.; Guo, B.; Tang, P.; and Liang, D. 2020. Non-Destructive Detection of Strawberry Quality Using Multi-Features of Hyperspectral Imaging and Multivariate Methods. *Sensors*, 20(11): 3074.
- [275] Wilder, B.; Dilkina, B.; and Tambe, M. 2019. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1658–1665.
- [276] Wojke, N.; Bewley, A.; and Paulus, D. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, volume 2017-Sept, 3645–3649.
- [277] Wu, M.-T.; and Tsai, C.-W. 2024. Training-free neural architecture search: A review. *ICT Express*, 10(1): 213–231.
- [278] Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2.
- [279] Xu, L.; and Zhao, Y. 2010. Automated strawberry grading system based on image processing. *Computers and Electronics in Agriculture*, 71(SUPPL. 1): S32–S39.

- [280] Xu, S.; Wang, J.; Wang, X.; and Shou, W. 2019. Computer Vision Techniques in Construction, Operation and Maintenance Phases of Civil Assets: A Critical Review. In *36th International Symposium on Automation and Robotics in Construction (ISARC 2019)*.
- [281] Xu, Y.; Imou, K.; Kaizu, Y.; and Saga, K. 2013. Two-stage approach for detecting slightly overlapping strawberries using HOG descriptor. *Biosystems Engineering*, 115(2): 144–153.
- [282] Xu, Y.; Zhou, X.; Chen, S.; and Li, F. 2019. Deep learning for multiple object tracking: a survey. *IET Computer Vision*, 13(4): 355–368.
- [283] Yamamoto, K.; Guo, W.; Yoshioka, Y.; and Ninomiya, S. 2014. On Plant Detection of Intact Tomato Fruits Using Image Analysis and Machine Learning Methods. *Sensors*, 14(7): 12191–12206.
- [284] Yan, R.; Wang, S.; and Fagerholt, K. 2020. A semi- “smart predict then optimize” (semi-SPO) method for efficient ship inspection. *Transportation Research Part B: Methodological*, 142: 100–125.
- [285] Yang, H.; Chang, F.; Huang, Y.; Xu, M.; Zhao, Y.; Ma, L.; and Su, H. 2022. Multi-object tracking using Deep SORT and modified CenterNet in cotton seedling counting. *Computers and Electronics in Agriculture*, 202: 107339.
- [286] Yang, L.; Fan, Y.; and Xu, N. 2019. Video Instance Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5187–5196.
- [287] Yang, Q.; and Jin, Z. 2024. Dataset for “Predicting the growth trajectory and yield of greenhouse strawberries based on knowledge-guided computer vision”. <https://doi.org/10.5281/zenodo.10957909>.
- [288] Yang, Q.; Simão, T. D.; Jansen, N.; Tindemans, S. H.; and Spaan, M. T. 2023. Reinforcement Learning by Guided Safe Exploration. *arXiv:2307.14316*.
- [289] Yang, X.; Fan, L.; Li, X.; and Meng, L. 2023. Day-ahead and real-time market bidding and scheduling strategy for wind power participation based on shared energy storage. *Electric Power Systems Research*, 214: 108903.
- [290] Yilmaz, A.; Javed, O.; and Shah, M. 2006. Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4): 13–es.
- [291] Young-Chang Chang; and Reid, J. 1996. RGB calibration for color image analysis in machine vision. *IEEE Transactions on Image Processing*, 5(10): 1414–1422.
- [292] Yu, Y.; Zhang, K.; Yang, L.; and Zhang, D. 2019. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Computers and Electronics in Agriculture*, 163: 104846.
- [293] Yue, X.-Q.; Shang, Z.-Y.; Yang, J.-Y.; Huang, L.; and Wang, Y.-Q. 2020. A smart data-driven rapid method to recognize the strawberry maturity. *Information Processing in Agriculture*, 7(4): 575–584.

- [294] Zhang, C.; Guo, C.; Liu, F.; Kong, W.; He, Y.; and Lou, B. 2016. Hyperspectral imaging analysis for ripeness evaluation of strawberry with support vector machine. *Journal of Food Engineering*, 179: 11–18.
- [295] Zhang, C.; Guo, C.; Liu, F.; Kong, W.; He, Y.; and Lou, B. 2016. Hyperspectral imaging analysis for ripeness evaluation of strawberry with support vector machine. *Journal of Food Engineering*, 179: 11–18.
- [296] Zhang, X.; Chowdhury, R. R.; Gupta, R. K.; and Shang, J. 2024. Large language models for time series: A survey. *arXiv preprint arXiv:2402.01801*.
- [297] Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; and Wang, X. 2022. ByteTrack: Multi-object Tracking by Associating Every Detection Box. In *European Conference on Computer Vision 2022*, 1–21. Springer, Cham.
- [298] Zhang, Z.; Zhong, B.; Zhang, S.; Tang, Z.; Liu, X.; and Zhang, Z. 2021. Distractor-Aware Fast Tracking via Dynamic Convolutions and MOT Philosophy. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1024–1033.
- [299] Zheng, C.; Abd-Elrahman, A.; and Whitaker, V. 2021. Remote sensing and machine learning in crop phenotyping and management, with an emphasis on applications in strawberry farming. *Remote Sensing*, 13(3): 531.
- [300] Zheng, Y.-Y.; Kong, J.-L.; Jin, X.-B.; Wang, X.-Y.; Su, T.-L.; and Zuo, M. 2019. CropDeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture. *Sensors*, 19(5): 1058.
- [301] Zhou, B.; Duan, X.; Ye, D.; Wei, W.; Woźniak, M.; Połap, D.; and Damaševičius, R. 2019. Multi-Level Features Extraction for Discontinuous Target Tracking in Remote Sensing Image Monitoring. *Sensors*, 19(22): 4855.
- [302] Zhou, Z.; Chen, J.; Pei, W.; Mao, K.; Wang, H.; and He, Z. 2022. Global Tracking via Ensemble of Local Trackers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8751–8760.
- [303] Zhou, Z.; Huang, Y.; Wang, W.; Wang, L.; and Tan, T. 2017. See the Forest for the Trees: Joint Spatial and Temporal Recurrent Neural Networks for Video-Based Person Re-identification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6776–6785.

# Acknowledgments

To finally write these words of gratitude has been one of the strongest motivations driving me toward the completion of this PhD. It feels deeply fitting that this dissertation should conclude not with data or arguments, but with a heartfelt thank you. As I sit down to write, countless names and faces come to mind. I have received so much love and support from them, and that sustains my optimism for work, life, and the world.

First and foremost, I am profoundly grateful for the opportunity to pursue this PhD, the true starting point of this entire journey. I owe my deepest thanks to Mathijs and Thomas, my supervisors, whose patience, encouragement, and unwavering belief in me have carried me through. Whenever I share stories about them with friends, especially with some PhD friends, the usual reaction is a mix of surprise (and envy). I always tell them, half-joking but fully sincere, that I've had the best team anyone could ask for.

I am also grateful for the crucial support from our project partners. I want to thank Birds.ai (Camiel, Irina, Jan Erik) and Delphy BV (Lisanne, Stijn, Klaas, Vera, and Brigit), whose help was vital, particularly at the beginning of the project. They not only helped me onboard quickly but also dedicated immense effort in data collection, forming the backbone of this research. My gratitude extends as well to FruitMasters and other project stakeholders for sharing insights from real-world practices and providing invaluable feedback on my solutions. I would like to acknowledge that this research was made possible by funding from Topsector Tuinbouw & Uitgangsmaterialen, Innovatiefonds Hagelunie, and Interpolis.

I began my PhD journey getting to know the Algorithmics and Bioinformatics groups as names and profile photos on Zoom and Mattermost. Thankfully, I eventually had the chance to connect with so many of you in person. I appreciate the time and interesting talks I shared with my office mates: Greg, Thiago, Qisong, Grisha, Leonard, Yaniv, Moritz, Pascal, Caroline, Simon, Liang, and Xueying. I'm also grateful to Canmanie, Clinton, Daniël, Eggie, Haohui, Imko, Issa, Kim, Konstantin, Koos, Ksenija, Laurens, Longjian, Maaïke, Maarten, Max, Noah, Reuben, Robert, Sterre, Tilman, and Yanyan for all the fun moments, shared food, inspiring and great conversations during group retreats, events, lunch, coffee breaks, and more time. Sometimes it felt like I was locked by my work when you invited, but I truly wish I could have joined more often. My thanks also go to Xucong and Kim for proofreading my work and offering suggestions that sharpened my ideas.

Two special thanks go to Qisong and Clinton for their consistent yet unique support. Thank you, Qisong, for your gentle but sharp reminder: "Do you really have so much spare time?" And thank you, Clinton, for the simple yet powerful advice: "Just accept what has happened and move on." Without your straightforward and unique warmth, which surprisingly works like a reset button for my emotions and reminds me you're always by my side, I might still be stuck in the whirlpool of my own thoughts, rather than standing here at the end of this long, improbable journey.

I moved to the Netherlands by myself in January 2021, in the midst of the pandemic. But thanks to the people I found here, I was never lonely. On my very first day, Kai Wu and Jingyi, who I met from a bachelor's exchange program, welcomed me with a fruitful and gigantic lunchbox. Soon after, Chinese New Year came. Qisong, who I only met once in our group meeting on Zoom, organized a gathering as both a new-year celebration and a housewarming party for me, bringing together Longjian, Jinke, Wei, and Yang. That was the beginning of our group, which grew from after-work dinners to weekend hangouts – If there's one period I will never forget, it's the two-month-long "farewell" with Bo, Jun Wen, Qisong, Ruoyun, Sen, Wei, Xueer, and Zilong. We cherished every moment as if it were our last time all together. And I promise, the dancing videos are safe with me! That group truly became my family here. But my community didn't stop there. I am so grateful for all the friends I have made in the Netherlands, whether through the help of these "social kings and queens" or just fun coincidences. I would like to thank Cheng, Chengmin, Chengyao, Chun-Ting, Dingyang, Guliuxin, Jing, Kai Wang, Li, Shiyu, Tingting, Wanru, Xiangwei, Xiaoyan, Xiaoyao, Xin, Xueting, Xusen, and Yucheng for all the great time and warm moments we shared. In particular, I would like to reserve my warmest thanks for Wei. You are the one that I will think of at the first place when I need a hug or when Diary and Luchi need an adult, and you never hesitate to accompany us. My life in Delft would have been much emptier without you.

Badminton has been a huge source of joy and friendship, with many of the people I met on the court becoming close friends beyond it. I am deeply grateful to my competition team members, my tournament partners, and all those I got to know through the games – Alizé, Anita, Ashlee, Bethany, Bofan, Chengji, Chi, Guan-Wei, Han, Hsiang-Jen, Jiahong, Jiani, Liang-Shin, Nancy, Qinyue, Tony, Tzu-Yueh, Xuesong, Yi-Kai, Yu, Yu-Chen, Yubo, Yvonne, and Zeyu. I especially cherish the laughter, encouragement, and memories we shared. Quite a few blocks on my calendar were marked with a competition or tournament name, which also led me to travel quite a lot and explore the surroundings, especially the smaller cities. We even made it to Germany for a group competition, which was a really fun trip! It may sound unrelated, but badminton taught me important qualities that carried me through this PhD. From my games and matches, I learned to be persevere and focus on giving my best and earning every point I could, no matter how strong the opponent, or how far behind I was. Along the way, I often found I was stronger than I thought – and those discoveries gave me the hope and drive to reach the finishing point of this journey.

My world has felt so much bigger and warmer, thanks to my friends scattered across nearly every time zone. Thank you to Ada, Arthur, Chikara, Chris, Claire, Danli, Darryl, Guoan, Haozhi, Isuru, Jingwei, Jingyuan, Jun Zhang, Kemeng, Lu, Luchi, Luqi, Maiya, Ningxing, Prakash, Sihan, Siqin, Wenjin, Xiaofan, Xiaohui, Xujia, Yanning, Yaran, Yi, Yihan, Yijia, Yueheng, Yueyang, and Yunwei for always welcoming me with open arms when I visited. Traveling and seeing you, whether for conferences or simply for fun, was always a reminder of the world beyond my PhD. Those visits gave me balance, perspective, and joy, and made each trip feel more special, more connected, and closer to life as a local. A huge thank you goes to Wenpei for designing the absolutely adorable cover for this thesis! I am truly glad to have all of you in my life. Your presence has filled this journey with warmth, friendship, and a sense of belonging that reached far beyond academia.



I want to give big thanks to my (furry) family here in the Netherlands: Diary, Luchi, and our temporary guest Schrödinger. Thank you for your tireless contributions: sitting on my keyboard, activating VS Code's autofill, squeezing past my monitor and showing up in my meetings from every angle, supervising my cooking (and occasionally hunting the ingredients), leaving evidences on my anonymized gifts, and diligently ensuring my life to be healthy by waking me up at early or ever-earlier hours in the morning. I would also like to thank you three for accompanying me at all time, no matter if I'm delighted, down, focused, or exhausted. I know I mean more than food and serves to you.

Finally, following a tradition of ending with the most important words, I give my deepest thanks to my parents and my grandma. Thank you for your unwavering support, for picking me up from the airport no matter how late the flight, for always having the simple, powerful words: that "home is always here". It gave me the courage to go so far and has meant more to me than I can say. I hope that now you can say, with pride and a smile: "Our daughter is a Doctor of Philosophy, and she has surpassed all those 'other kids'!"

And a final, quiet thank you to Junhan, myself. For the motivation, the persistence, and all the other complimentary words I could list... but really, for simply not giving up. This thesis marks more than four years of research; it is also a reflection of the life that unfolded alongside it. The work and the living were never separate, and together they shaped not only my résumé but also my way of seeing and being. The year I completed this work is also the year I turned thirty, a milestone not only in my academic journey but in my life. These years have left their mark upon me, and I look forward to who I will be.

*Junhan Wen*  
文君涵  
*Delft, 2025.10.01*





# Curriculum Vitæ

## Junhan Wen

1995/07/07      Born in Wuhan, Hubei, China

### Education






- 2021/01 - 2025/10      Ph.D. in Computer Science,  
Technische Universiteit Delft (TU Delft), Delft, Netherlands.  
Project: *An Optimal Soft-Fruit Chain Starts with the Harvest*  
Promotors: Prof. Dr. Mathijs de Weerd, Dr. Thomas Abeel
- 2017/09 - 2021/01      M.Sc. in Integrated Building Systems,  
Eidgenössische Technische Hochschule Zürich (ETH Zurich),  
Zurich, Switzerland.  
Thesis: *Scheduling Production Capacities with Data-Driven Demand Forecasting*
- 2013/09 - 2017/07      B.Eng. in Civil Engineering,  
Southeast University, Nanjing, China.  
Thesis: *Upper Structural Design and Mechanical Analysis of the Dike-Crossing Bridge of the 5<sup>th</sup> Nanjing Yangtze River Bridge*


### Experience


- 2025/01 - 2025/08      Researcher,  
Honda Research Institute Co. Ltd. Japan
- 2020/09 - 2021/01      Research Assistant,  
Dept. of Humanities, Social and Political Sciences, ETH Zurich
- 2019/06 - 2019/12      Research Assistant,  
Dept. of IT for Innovative Service, Luxembourg Institute of  
Science and Technology



# List of Publications

10. *Junhan Wen*, X. Zhang, J. Chew. “Anticipating Daily Human Actions: Comparing Pipelines for Long-Term Skeleton-Based Prediction in Real-World Scenarios”. *under review of Advanced Robotics*.
-  9. *Junhan Wen*, T. Abeel, M. de Weerd. “Data-Driven Precision Harvest Planning via Visual Growth Forecasting”. *under review of Applied Artificial Intelligence*.
-  8. *Junhan Wen*, T. Abeel, M. de Weerd. 2025. “Performance and Interaction Assessment of Neural Network Architectures and Bivariate Smart Predict-then-Optimize”. *Machine Learning Journal (MLJ)*.
-  7. *Junhan Wen*, C. R. Verschoor, C. Feng, I. Epure, T. Abeel, M. de Weerd. 2024. “The Growing Strawberries Dataset: Tracking Multiple Objects With Biological Development Over an Extended Period”. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
-  6. *Junhan Wen*, A. Dascalu, D. Drashkov, L. Marinski, M. A. Bhatti, M. Herrebut, M. Weinmann. 2023. “Demonstration of a Mobile Application for Near Real-Time Strawberry Quality Prediction”. *BNAIC/BeNeLearn*.
5. C. Jol, *Junhan Wen*, J. van Gemert. 2023. “Non-Destructive Infield Quality Estimation of Strawberries using Deep Architectures”. *ICCV Computer Vision in Plant Phenotyping and Agriculture*.
-  4. *Junhan Wen*, T. Abeel, M. de Weerd. 2023. “‘How Sweet Are Your Strawberries?’: Predicting Sugariness using Non-Destructive and Affordable Hardware”. *Frontiers in Plant Science*.
3. *Junhan Wen*, T. Abeel, M. de Weerd. 2021. “Towards Predicting the Maturity and Quality of Strawberries in Greenhouses”. *AAAI 2021 Fall Symposium Series*.
2. *Junhan Wen*, E. Daher, S. Kubicki. 2020. “Optimization of a User-Involved Floor Layout Recommendation System at the Operation Stage”. *CIB W78*.
1. B. Gahr, K. Caves, *Junhan Wen*, K. Koch, S. Liu, and F. Wortmann. 2019. “The Costs of Traffic Accident Hotspots”. *IEEE ITSC*.

 Included in this thesis.

 Won a best demonstration award.

# Titles in the SIKS Dissertation Series since 2016

- 
- 2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
  - 02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
  - 03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
  - 04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
  - 05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
  - 06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
  - 07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
  - 08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
  - 09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
  - 10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
  - 11 Anne Schuth (UvA), Search Engines that Learn from Their Users
  - 12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
  - 13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
  - 14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
  - 15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
  - 16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
  - 17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
  - 18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
  - 19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
  - 20 Daan Odijk (UvA), Context & Semantics in News & Web Search
  - 21 Alejandro Moreno Céleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
  - 22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems

- 23 Fei Cai (UvA), Query Auto Completion in Information Retrieval
- 24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
- 25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
- 26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
- 27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
- 28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
- 29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
- 30 Ruud Mattheij (TiU), The Eyes Have It
- 31 Mohammad Khelghati (UT), Deep web content monitoring
- 32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
- 33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
- 34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
- 35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
- 36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
- 37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
- 38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
- 39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
- 40 Christian Detweiler (TUD), Accounting for Values in Design
- 41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
- 42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
- 43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
- 44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
- 45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
- 46 Jorge Gallego Perez (UT), Robots to Make you Happy
- 47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
- 48 Tanja Buttler (TUD), Collecting Lessons Learned

- 
- 49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
- 50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains
- 
- 2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
- 02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
- 03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
- 04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
- 05 Mahdieh Shadi (UvA), Collaboration Behavior
- 06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
- 07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
- 08 Rob Konijn (VUA), Detecting Interesting Differences: Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
- 09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
- 10 Robby van Delden (UT), (Steering) Interactive Play Behavior
- 11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
- 12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
- 13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
- 14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
- 15 Peter Berck (RUN), Memory-Based Text Correction
- 16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
- 17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
- 18 Ridho Reinanda (UvA), Entity Associations for Search
- 19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
- 20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
- 21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
- 22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
- 23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces
- 24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
- 25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
- 26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
- 27 Michiel Jooisse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors



- 
- 28 John Klein (VUA), Architecture Practices for Complex Contexts
  - 29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
  - 30 Wilma Latuny (TiU), The Power of Facial Expressions
  - 31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
  - 32 Thaer Samar (RUN), Access to and Retrieval of Content in Web Archives
  - 33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
  - 34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
  - 35 Martine de Vos (VUA), Interpreting natural science spreadsheets
  - 36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
  - 37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
  - 38 Alex Kayal (TUD), Normative Social Applications
  - 39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
  - 40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
  - 41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
  - 42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
  - 43 Maaïke de Boer (RUN), Semantic Mapping in Video Retrieval
  - 44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
  - 45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
  - 46 Jan Schneider (OU), Sensor-based Learning Support
  - 47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
  - 48 Angel Suarez (OU), Collaborative inquiry-based learning
- 
- 2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
  - 02 Felix Mannhardt (TU/e), Multi-perspective Process Mining
  - 03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
  - 04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
  - 05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
  - 06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
  - 07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
  - 08 Rick Smetsters (RUN), Advances in Model Learning for Software Systems

- 09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
  - 10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
  - 11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
  - 12 Xixi Lu (TU/e), Using behavioral context in process mining
  - 13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
  - 14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
  - 15 Naser Davarzani (UM), Biomarker discovery in heart failure
  - 16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
  - 17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
  - 18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
  - 19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
  - 20 Manxia Liu (RUN), Time and Bayesian Networks
  - 21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
  - 22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
  - 23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
  - 24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
  - 25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
  - 26 Roelof Anne Jelle de Vries (UT), Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
  - 27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
  - 28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
  - 29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech
  - 30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web
- 
- 2019 01 Rob van Eijk (UL), Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
  - 02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
  - 03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
  - 04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
  - 05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
  - 06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
  - 07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
  - 08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

- 09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems
- 10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
- 11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
- 12 Jacqueline Heinerman (VUA), Better Together
- 13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
- 14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
- 15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
- 16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
- 17 Ali Hurriyetoglu (RUN), Extracting actionable information from microtexts
- 18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
- 19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
- 20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
- 21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
- 22 Martin van den Berg (VUA), Improving IT Decisions with Enterprise Architecture
- 23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
- 24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
- 25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
- 26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
- 27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
- 28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
- 29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
- 30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
- 31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
- 32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
- 33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
- 34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

- 
- 35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
  - 36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
  - 37 Jian Fang (TUD), Database Acceleration on FPGAs
  - 38 Akos Kadar (OU), Learning visually grounded and multilingual representations
- 
- 2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
  - 02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
  - 03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
  - 04 Maarten van Gompel (RUN), Context as Linguistic Bridges
  - 05 Yulong Pei (TU/e), On local and global structure mining
  - 06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
  - 07 Wim van der Vegt (OU), Towards a software architecture for reusable game components
  - 08 Ali Mirsoleimani (UL), Structured Parallel Programming for Monte Carlo Tree Search
  - 09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
  - 10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
  - 11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation Methods for Long-Tail Entity Recognition Models
  - 12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
  - 13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
  - 14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
  - 15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
  - 16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
  - 17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
  - 18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
  - 19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
  - 20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
  - 21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
  - 22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
  - 23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
  - 24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots

- 
- 25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
  - 26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
  - 27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
  - 28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
  - 29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
  - 30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
  - 31 Gongjin Lan (VUA), Learning better – From Baby to Better
  - 32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
  - 33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
  - 34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
  - 35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production
- 
- 2021 01 Francisco Xavier Dos Santos Fonseca (TUD), Location-based Games for Social Interaction in Public Space
  - 02 Rijk Mercur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
  - 03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
  - 04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
  - 05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
  - 06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
  - 07 Armel Lefebvre (UU), Research data management for open science
  - 08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
  - 09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
  - 10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
  - 11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
  - 12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
  - 13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
  - 14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
  - 15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
  - 16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
  - 17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks

- 
- 18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
  - 19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
  - 20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
  - 21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
  - 22 Sihang Qiu (TUD), Conversational Crowdsourcing
  - 23 Hugo Manuel Proença (UL), Robust rules for prediction and description
  - 24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
  - 25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
  - 26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
  - 27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
  - 28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs
- 
- 2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
  - 02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
  - 03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
  - 04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
  - 05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
  - 06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
  - 07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
  - 08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
  - 09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
  - 10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
  - 11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
  - 12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
  - 13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
  - 14 Michiel Overeem (UU), Evolution of Low-Code Platforms

- 15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
  - 16 Pieter Gijbbers (TU/e), Systems for AutoML Research
  - 17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
  - 18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
  - 19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
  - 20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
  - 21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
  - 22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
  - 23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
  - 24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
  - 25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
  - 26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
  - 27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
  - 28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
  - 29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
  - 30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
  - 31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
  - 32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
  - 33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
  - 34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
  - 35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction
- 
- 2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
  - 02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
  - 03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
  - 04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval



- 
- 05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
  - 06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
  - 07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
  - 08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
  - 09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
  - 10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
  - 11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
  - 12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
  - 13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
  - 14 Selma Čaušević (TUD), Energy resilience through self-organization
  - 15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
  - 16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
  - 17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision –From Oversight to Insight
  - 18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
  - 19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
  - 20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
  - 21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
  - 22 Alireza Shojaiifar (UU), Volitional Cybersecurity
  - 23 Theo Theunissen (UU), Documentation in Continuous Software Development
  - 24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
  - 25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
  - 26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
  - 27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
  - 28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
  - 29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results
-



- 2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
- 02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
- 03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
- 04 Mike Huisman (UL), Understanding Deep Meta-Learning
- 05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
- 06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
- 07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
- 08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
- 09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
- 10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
- 11 withdrawn
- 12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
- 13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
- 14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
- 15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
- 16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
- 17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
- 18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
- 19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
- 20 Ritsart Anne Plantenga (UL), Omgang met Regels
- 21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
- 22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
- 23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
- 24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
- 25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
- 26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
- 27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning

- 
- 28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs
  - 29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
  - 30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
  - 31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
  - 32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
  - 33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction
  - 34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
  - 35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
  - 36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
  - 37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
  - 38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
  - 39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
  - 40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
  - 41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
  - 42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
  - 43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
  - 44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
  - 45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
  - 46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
  - 47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
  - 48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions
- 
- 2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
  - 02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
  - 03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI
  - 04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data

- 05 Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation
- 06 Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
- 07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
- 08 Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
- 09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach
- 10 Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
- 11 Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
- 12 Lijun Lyu (TUD), Interpretability in Neural Information Retrieval
- 13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
- 14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies
- 15 Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence
- 16 Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging
- 17 Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems
- 18 Anouk Neerinx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing
- 19 Fang Hou (UU), Trust in Software Ecosystems
- 20 Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)
- 21 Mandani Ntekouli (UM), Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data
- 22 Hilde Weerts (TU/e), Decoding Algorithmic Fairness: Towards Interdisciplinary Understanding of Fairness and Discrimination in Algorithmic Decision-Making
- 23 Roderick van der Weerd (VUA), IoT Measurement Knowledge Graphs: Constructing, Working and Learning with IoT Measurement Data as a Knowledge Graph
- 24 Zhong Li (UL), Trustworthy Anomaly Detection for Smart Manufacturing
- 25 Kyana van Eijndhoven (TiU), A Breakdown of Breakdowns: Multi-Level Team Coordination Dynamics under Stressful Conditions
- 26 Tom Pepels (UM), Monte-Carlo Tree Search is Work in Progress
- 27 Danil Provodin (JADS, TU/e), Sequential Decision Making Under Complex Feedback

- 28 Jinke He (TU Delft), Exploring Learned Abstract Models for Efficient Planning and Learning
- 29 Erik van Haeringen (VUA), Mixed Feelings: Simulating Emotion Contagion in Groups
- 30 Myrthe Reuver (VUA), A Puzzle of Perspectives: Interdisciplinary Language Technology for Responsible News Recommendation
- 31 Gebrekirstos Gebreselassie Gebremeskel (RUN), Spotlight on Recommender Systems: Contributions to Selected Components in the Recommendation Pipeline
- 32 Ryan Brate (UU), Words Matter: A Computational Toolkit for Charged Terms
- 33 Merle Reimann (VUA), Speaking the Same Language: Spoken Capability Communication in Human-Agent and Human-Robot Interaction
- 34 Eduard C. Groen (UU), Crowd-Based Requirements Engineering
- 35 Urja Khurana (VUA), From Concept To Impact: Toward More Robust Language Model Deployment
- 36 Anna Maria Wegmann (UU), Say the Same but Differently: Computational Approaches to Stylistic Variation and Paraphrasing
- 37 Chris Kamphuis (RUN), Exploring Relations and Graphs for Information Retrieval
- 38 Valentina Maccatrozzo (VUA), Break the Bubble: Semantic Patterns for Serendipity
- 39 Dimitrios Alivanistos (VUA), Knowledge Graphs & Transformers for Hypothesis Generation: Accelerating Scientific Discovery in the Era of Artificial Intelligence
- 40 Stefan Grafberger (UvA), Declarative Machine Learning Pipeline Management via Logical Query Plans
- 41 Mozghan Vazifehdoostirani (TU/e), Leveraging Process Flexibility to Improve Process Outcome - From Descriptive Analytics to Actionable Insights
- 42 Margherita Martorana (VUA), Semantic Interpretation of Dataless Tables: a metadata-driven approach for findable, accessible, interoperable and reusable restricted access data
- 43 Krist Shingjergji (OU), Sense the Classroom - Using AI to Detect and Respond to Learning-Centered Affective States in Online Education
- 44 Robbert Reijnen (TU/e), Dynamic Algorithm Configuration for Machine Scheduling Using Deep Reinforcement Learning
- 45 Anjana Mohandas Sheeladevi (VUA), Occupant-Centric Energy Management: Balancing Privacy, Well-being and Sustainability in Smart Buildings
- 46 Ya Song (TU/e), Graph Neural Networks for Modeling Temporal and Spatial Dimensions in Industrial Decision-making
- 47 Tom Kouwenhoven (UL), Collaborative Meaning-Making. The Emergence of Novel Languages in Humans, Machines, and Human-Machine Interactions
- 48 Evy van Weelden (TiU), Integrating Virtual Reality and Neurophysiology in Flight Training
- 49 Selene Báez Santamaría (VUA), Knowledge-centered conversational agents with a drive to learn

- 
- 50 Lea Krause (VUA), Contextualising Conversational AI
  - 51 Jiaxu Zhao (TU/e), Understanding and Mitigating Unwanted Biases in Generative Language Models
  - 52 Qiao Xiao (TU/e), Model, Data and Communication Sparsity for Efficient Training of Neural Networks
  - 53 Gaole He (TUD), Towards Effective Human-AI Collaboration: Promoting Appropriate Reliance on AI Systems
  - 54 Go Sugimoto (VUA), MISSING LINKS Investigating the Quality of Linked Data and its Tools in Cultural Heritage and Digital Humanities
  - 55 Sietze Kai Kuilman (TUD), AI that Glitters is Not Gold: Requirements for Meaningful Control of AI Systems
  - 56 Wijnand van Woerkom (UU), A Fortiori Case-Based Reasoning: Formal Studies with Applications in Artificial Intelligence and Law
  - 57 Syeda Amna Sohail (UT), Privacy-Utility Trade-Off in Healthcare Metadata Sharing and Beyond: A Normative and Empirical Evaluation at Inter and Intra Organizational Levels
  - 58 Junhan Wen (TUD), “From iMage to Market”: Machine-Learning-Empowered Fruit Supply
- 
- 2026 01 Pei-Yu Chen (TUD), Human-Agent Alignment Dialogues: Eliciting User Information at Runtime for Personalized Behavior Support
-

