HORIZON EUROPE
Digital and emerging technologies for competitiveness and fit for the Green Deal

HYPERIMAGE

A universal spectral imaging sensor platform for industry, agriculture, and autonomous driving.

Starting date of the project: 01/12/2023
Duration: 42 months

= Deliverable D4.2 =

Ground truth data and digital twins of products

Due date of deliverable: 30/06/2025
Actual submission date: 02/07/2025

Responsible WP: Netcompany (INTRA), WP4
Responsible TL: Joseph Peller, WUR
Version: V1.0

Dissemination level		
PU	Public	x
SE	SENSITIVE – limited under the conditions of the Grant Agreement	

This project has received funding from the European Union's Horizon Program under grant agreement No.101135827

AUTHOR

Author	Institution	Contact (e-mail, phone)
Joseph Peller	WUR	Joseph.Peller@wur.nl, +31628486341
John Fahltech	Ketmarket	john.fahlteich@ketmarket.eu, +4915140707380

DOCUMENT CONTROL

Document version	Date	Change
V0.1	05/06/2025	Initial draft
V0.3	07/06/2025	Added information on dataset structures
V0.4	20/06/2025	Added information about NEO dataset
V1.0	01/07/2025	Final Version

VALIDATION

Reviewers		Validation date
Work Package Leader	Netcompany (INTRA)	01/07/2025
Project Manager	Marina de Souza Faria	01/07/2025
Project Coordinator	Alexander Kabardiadi-Virkovski	02/07/2025

DOCUMENT DATA

Keywords	HSI, spectral image, data format
Point of Contact	Name: Joseph Peller Partner: WUR Address: Hoge Steeg 2 Building 105 6708 PH Wageningen E-mail: joseph.peller@wur.nl
Delivery date	02/07/2025

DISCLAIMER

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Executive Summary

This document provides an overview of the data sets uploaded by HyperImage partners to serve as ground truth data and historical data for the cloud platform.

The four datasets are comprised of historical data from prior experiments and trials from the partners. Where each industrial partner did not have historical data, a knowledge institute supplied similar data that could be used to test the algorithms. These datasets were uniformly labeled and the structure determined in D4.1 was applied.

Finally, the GDPR and Economic implications of each dataset were considered.

Table of Contents

Table of Figures

1. Introduction

1.1. T4.2 Description from Grant Agreement

T4.2: Ground truth data and digital twins of products (WUR, FHG-IWS, use case partners | M9-M18)

4.2 collect reference image data related to the four industrial use cases described in T1.2 with state-of-the-art spectral imaging systems by WUR and FHG-IWS. Each use case partner (GROWY, NEO, ROB, IFBIP) will contribute by providing reference materials or image datasets.

The collected data sets will be analysed based on the goals of each use case (e.g., object detection, defects identification, calculation of derived properties on the observed goods, etc.) and labelled so that they will establish the ground truth for the training of the Machine Learning (ML)-based image analysis algorithms (FHG-IWS, WUR) (T4.3).The spectral imaging data, the associated ground truth values and the meta data are processed according to the data structure defined in T4.1.

The task will also specify the access control rights and any relevant technical, legal (GDPR) and commercial aspects of the sharing of this data to the consortium (and to external users) beyond the duration of the projects

1.2. Objectives of the deliverable

- Collection of reference image data related to the 4 industrial use cases
- Establish goals of each use case, and labelling to establish the ground truth data
- Perform data analysis on collected data to establish a baseline probability of success (no ground truth data)
- Specification of data access control rights and technical, legal (GDPR), commercial aspects related to data sharing

2. Collection of Ground Truth datasets

To properly test the and calibrate the algorithms that will be developed within the HyperImage project representative datasets of each of the partners needed to be produced and shared between all members of the consortium. Each dataset would consist of historical data generated by each industrial partner and would be reprocessed and formatted into a format which allowed their use in the cloud platform. Finally, the Commercial and privacy concerns of the data stored on the platform where considered.

2.1. Collection of reference image data

The four industrial partners contributed each dataset from prior projects that could be used as ground truth data for testing the algorithms within the cloud platform. Each Partner GROWY, DIVE, NEO, and Robotnik contributed a representative dataset, which shared similarities to the upcoming use cases. GROWY and WUR contributed a dataset of top-down spectral images of Lettuce, DIVE contributed images from wafer production, NEO contributed scenes of vegetation and geological outcrops from drone images, and Robotnik and SILIOS contributed a dataset of spectral images of various off-road scenes.

Each partner contributed a different number of datasets, GROWY and WUR contributed 31, DIVE contributed 41, NEO contributed 3 very large mosaic images, and Robotnik and SILIOS contributed 206 images. Each Image was saved using the folder structure laid out in task 4.1

2.2. Establishment of labeling and goals of each use case

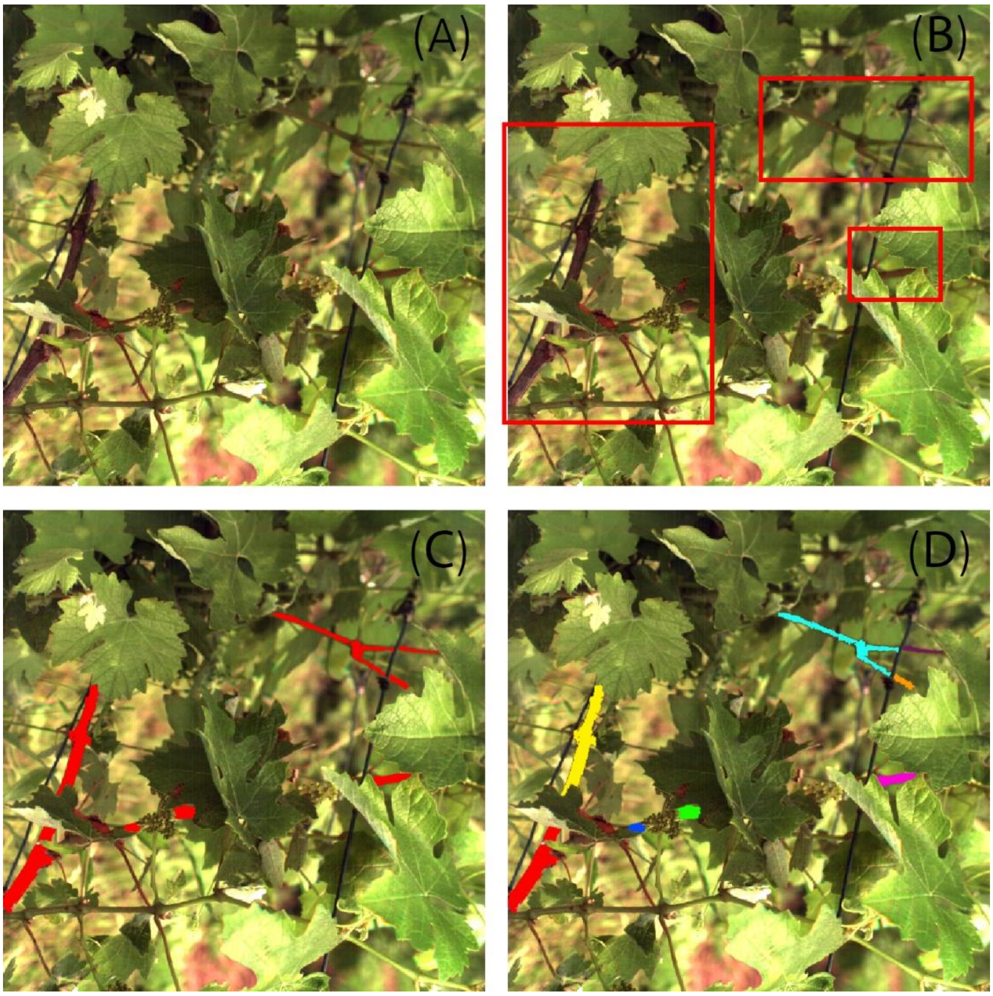

Figure 1: Examples of the three types of image labelling. In A) we have an image of grape leaves, and we want to find all the woody branches. B) shows an object detector, here the locations of the branches are found and closed in a bounding box. C) Shows Semantic Segmentation, where all pixels identified as a "woody branch" are labelled in the image, and in D) we have an instance segmentation where corresponding "woody branch" pixels form larger objects in the image.

All images need to provide ground truth data, this is done by *labelling* the images, or selecting what parts of the image are useful to algorithms. The simplest type of labelling is *object detection,* which simply identifies and locates objects within an image using bounding boxes. Moving towards more complex types, *semantic segmentation* classifies each pixel of an image into a category but makes no distinction between different objects of the same class. Finally, *instance segmentation* not only categorizing pixels but also identifying each distinct object of the same type.Each Dataset needed to be labelled according to the goals of each use case. For the GROWY Use case it is most important to segment out of each image the vegetation so that spectral analysis can be performed later. For Dive, segmentation of the chip bed is needed for wafer quality determination, and for Robotnik, semantic segmentation of different objects in the environment such as trees and obstacles. For NEO semantic segmentation was also done using spectral indices such as NDVI.

These different use cases therefore require different labelling methods. For both the GROWY and Dive cases instance segmentation was performed creating binary masks using spectral information to segment the objects from the background. For the Robotnik case eventually a neural network will run on the data to provide bounding boxes of each obstacle in the field, these labels will be stored in the JSON format.

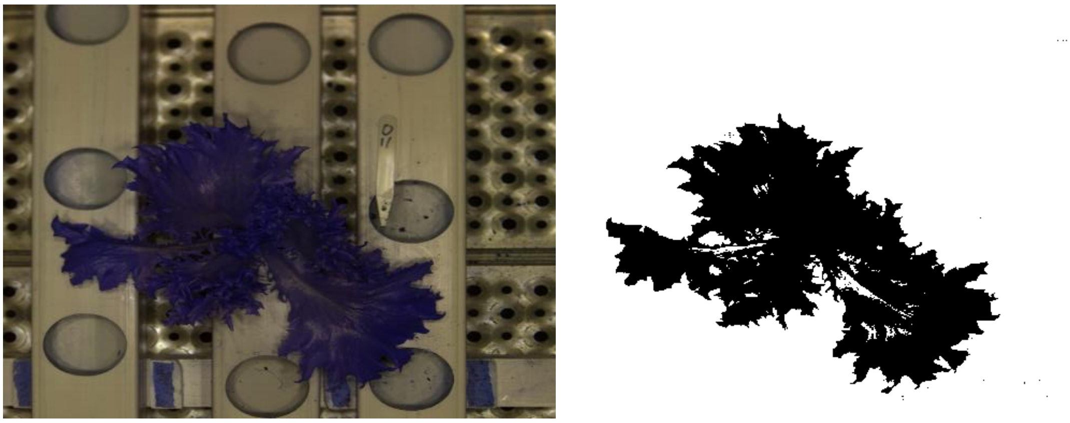

Figure 2: Example of instance segmentation by the GROWY use case. A small lettuce plant was segmented from the background using its spectral information creating a binary mask.

2.3. Data analysis on collected data to establish baseline probability of success

A digital twin is a virtual model designed to accurately reflect a physical object or system. It integrates IoT sensors and other data sources to create a real-time simulation that can predict behaviour, optimize performance, and guide decision-making processes. This technology finds application across various industries, from agriculture to semiconductor manufacturing, showcasing its versatility and utility.

In agriculture, a digital twin might simulate a farming environment, including crop growth, soil conditions, and weather patterns, to help growers optimize crop yield. Or as in the semiconductor industry, digital twins are used to model production processes. This includes everything from the assembly line workflows to the microscopic interactions during chip fabrication, enabling engineers to predict failures and optimize chip yields without the costly trial-and-error of physical testing.

These algorithms are central to how digital twins function and may require extensive data for training to achieve high accuracy and utility. The amount and type of data required can vary based on several factors. Complex systems, such as those with multiple variables or AI, require more robust datasets to train algorithms effectively. But simpler systems that rely on machine learning and mathematical correlations can require much less. The level of precision needed in each use case also influences the how varied the training data needs to be, but in general higher precision in the algorithms require richer data sets.

2.4. Specification of data access control rights and technical, legal (GDPR), commercial aspects related to the data sharing

In response to increasing concerns over data privacy and compliance, our project has implemented robust access control measures to ensure that data from varied domains such as self-driving cars, silicon wafer production, drone imaging, and vertical farming is handled in strict accordance with the General Data Protection Regulation (GDPR).

Currently access to the data is restricted by a password protected cloud server hosted in Hetzner Cloud, a public cloud provider. Password and user access is strictly controlled by Netcompany (INTRA) that has acquired the infrastructure, and the historical data is not available to the general public in its raw form. It will be used to train the algorithms used in the project, and those weights will be made open to commercial exploitation.

To align our operations with GDPR, we have adopted the best practices for prioritizing user privacy. This includes employing techniques for anonymizing and pseudonymizing personal data, thus significantly lowering the risk associated with data handling and storage. The cloud platform will allow individuals to access, rectify, delete, or object to the processing of their data by the cloud platform. Ensuring that these methods are in place early on helps in maintaining the homogeneity of data formats across disparate datasets but also in reinforcing data security upon the release of the platform.

3. Results and Discussion

We have successfully shown in this deliverable the ability to store and format spectral data from a wide variety of use cases and store them online for use in the HyperImage cloud platform.

One of the most important successes has been the application of the data structures specified in deliverable 4.1 and testing its applicability to the use cases. By harmonizing the images from the datasets, we have created a methodology that is both flexible and useful in each of our fields. Harmonization across the spectral imaging field is important, and it is important that HyperImage does not create yet another standard in a sea of standards. Instead, we have offered a container structure for the existing ENVI format that allows for users to quickly access, preview data as well as label it in multiple fashions.

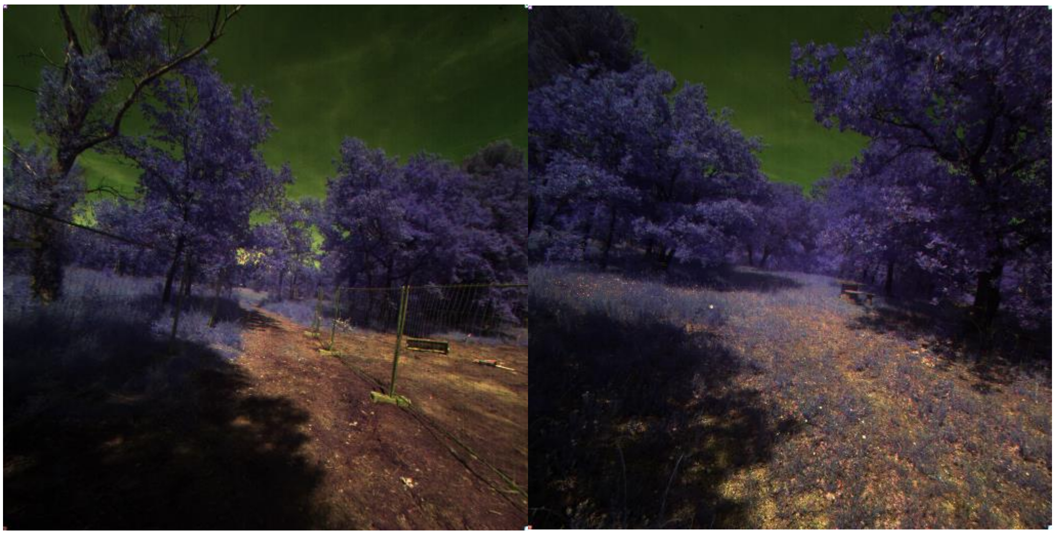

Figure 3: Example of false color RGBs for the Robotnik outdoor vehicle case. Showing the enhanced Contrast between Vegetation sky and ground.

4. Conclusions

This deliverable has successfully met the objectives outlined in the initial stages of the project. We collected and processed spectral imaging data while meeting the requirements from four distinct industrial use cases, which were described in the grant agreement. The datasets from all four collaborating partners: GROWY, NEO, ROB, and Dive, have been collected and labeled to serve as a ground truth for machine learning algorithms dedicated to image analysis described in the rest of the Work Package 4. This milestone sets a foundation for the ongoing development and refinement of predictive models for the rest of the project.

5. Degree of Progress

The four datasets of historical data have been uploaded, labeled, and structured in a uniform way. This data can now be used to test algorithms and become a baseline for developments in WP6 and WP7. All of the requirements for this have been fully met including considering the privacy and legal obligations of using such data.

6. Dissemination Level

Public.