Eindhoven University of Technology

Department of Mathematics and Computer Science

In collaboration with VBTI Consultancy B.V.

Semi-Supervised Learning Optimization for Industrial AI Deployment:

A Pseudo-Labeling Framework for Cost-Optimized Computer Vision Pipeline Development

Bachelor Thesis
Data Science

Daniel Osman

Student ID: 1818783

Supervised by,

Illya Kaynov
Dieter Timmers
Maryam Tavakol

July 2025

Contents

Abstract

The success and performance of deep learning models in computer vision tasks heavily depend on the availability of high-quality annotated training data. Despite the importance of annotated data, the process of manually annotating images is both extremely time-intensive and costly, and is traditionally outsourced to external annotators. This process is especially challenging and expensive for complex tasks like instance segmentation with multiple classes and instances. This research investigates the effectiveness of implementing a semi-supervised learning approach using pseudo-labeling techniques to streamline the annotation process while maintaining competitive model performance and reducing costs. This research analyzes the strengths of using a pseudo-labeling pipeline within complex agricultural environments and develops a comprehensive pseudo-labeling structure that iteratively improves model accuracy by using an initial model trained on a small subset of manually annotated data to generate labels on unseen images then retrain. The framework was evaluated on two complex real-world agricultural datasets. The first one uses an asparagus dataset for object detection (12,630 images with 5 classes) and the second uses a cucumber dataset for instance segmentation (2,429 images with 7 classes averaging 70 instances per image). Experiments were conducted separately on each dataset using six different pseudo-labeling strategies tested across five iterations, all with varying balance between ground truth annotations and pseudo-labels. The goal was to determine the effectiveness of pseudo labeling in real world agricultural scenarios and establish optimal supervision thresholds for each use case. The results from the research reveal that pseudo-labeling can achieve substantial cost reductions while maintaining competitive performance. For object detection, the optimal flow (F3) achieved 86.4% of baseline performance while requiring only 43.75% manual annotation effort (56.3% cost reduction). For instance segmentation, flow F4-S achieved 100.5% performance retention with a 75% cost reduction. The segmentation task showed consistently higher efficiency ratios (3.62-43.84) compared to object detection (3.07-5.51), which highlights the value of pseudo-labeling for dense annotation tasks. Manual correction effort decreased by 57.4% across iterations which further illustrates the self-improving nature of the pipeline after each iteration. Beyond the experiments conducted, this research also produced a pseudo-labeling framework integrated within VBTI's OneDL platform. This framework allows employees and clients of VBTI to implement pseudo-labeling across any computer vision project with features like automated database logging, integration with annotation platforms for manual corrections, and support for multiple model architectures. The framework has already been adapted to one of the projects, and several colleagues have shown interest in using the framework for their own real-world applications. VBTI has also shown interest and is considering integrating the framework into their front-end platform. The findings in this paper demonstrate that pseudo-labeling is an effective strategy for reducing annotation costs and accelerating proof-of-concept development in industrial AI applications, specifically for agricultural computer vision tasks.