Industrial Part Counting using Computer Vision

Master's thesis report in Production Engineering, MSc

MAJA ADLER, SANDRA MELANDER

DEPARTMENT OF INDUSTRIAL AND MATERIALS SCIENCE

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

Industrial Part Counting using Computer Vision

MAJA ADLER, SANDRA MELANDER

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Industrial and Materials Science
Chalmers University of Technology
Gothenburg, Sweden 2026

Industrial Part Counting using Computer Vision
MAJA ADLER, SANDRA MELANDER

Industrial Supervisor: Ian Rathgeb, Senior Logistics Engineer at Volvo Trucks Tuve.
Supervisor: Hao Wang, Department of Industrial and Materials Science.
Examiner: Björn Johansson, Department of Industrial and Materials Science.

Master's Thesis 2026
Department of Industrial and Materials Science
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Gothenburg, Sweden 2026

Industrial Part Counting using Computer Vision
MAJA ADLER, SANDRA MELANDER
Department of Industrial and Materials Science
Chalmers University of Technology

Abstract

Reliable quantity verification is important in logistics and manufacturing, where incorrect deliveries can lead to delays, customer claims, and increased costs. At the complete knock down (CKD) operations at Volvo Trucks Tuve, quantity verification is currently performed manually, making the process time-consuming and vulnerable to human error. Therefore, this thesis investigates the potential of using computer vision for industrial part counting.

The aim of the thesis was to evaluate computer vision-based part counting at CKD packing stations by reviewing the current state of the art and assessing the performance of the Volvo Vision System (VVS). The study followed a design science research methodology approach and included a current state analysis, stakeholder analysis, interviews, observations, a literature review, and practical tests using VVS and YOLOv5 object detection models. Different hardware settings, lighting conditions, and dataset sizes were tested to evaluate the feasibility of the system in an industrial environment.

The results show that computer vision-based object counting has potential for industrial applications, primary in structured environments. The experiments demonstrated that VVS could successfully detect and count several industrial parts under controlled conditions. Increasing the number of training images per class slightly improved the model performance, and the system achieved high accuracy for multiple object classes. However, the study also identified several challenges related to lighting variations, object overlap, reflections, scalability, and long-term maintainability.

The study concludes that VVS has high potential and performance when it comes to object detection. However, in cases such as this one where there is a large amount of highly varied products and manual work, there are a lot of challenges that hinders a smooth implementation of a vision system.

Keywords: Computer vision, Object counting, Part counting, YOLOv5, Industry, Logistics, Machine learning, Object detection.

Acknowledgements

We would like to express our sincere gratitude to everyone who supported us throughout this project and made our time at Volvo Tuve so rewarding.

First, we would like to thank our examiner, Björn Johansson, and our supervisor, Hao Wang, for all the guidance, support, and encouragement during the entire project. Thank you for always being easy to reach, for your valuable feedback and ideas, and for helping us shape both our work and report in the best possible way.

We also want to give a special thank you to our industrial supervisor, Ian Rathgeb, who showed genuine interest in our work from day one. Thank you for introducing us to the right people, taking the time to explain processes, showing us around, and always being willing to help whenever we had questions.

We are also very grateful to the entire team at the CKD operations, especially the team leaders and all operators, for always welcoming us with open minds and being so willing to discuss, share insights, and help us throughout the project.

Finally, we would like to thank the VVS team for helping us understand the VVS processes and for always taking the time to answer our questions and support us.

Lastly, we would also like to thank everyone else at Volvo Tuve who contributed to making this experience both educational and enjoyable for us.

Maja Adler & Sandra Melander, Gothenburg, June 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

AI	Artificial intelligence
CMOS	Complementary metal oxide semiconductor
CNN	Convolutional neural network
CKD	Complete Knock Down
DL	Deep learning
DSR	Design science research
DSRM	Design science research methodology
FOV	Field of view
IoT	Internet of Things
LVLM	Large Vision-Language Model
ML	Machine learning
RCA	Root cause analysis
RGB	Red, green, blue
UCR	Unpacking Control Reports
UI	User interface
VVS	Volvo Vision System
YOLO	You Only Look Once

Nomenclature

Below is the nomenclature of indices, sets, parameters, and variables that have been used throughout this thesis.

Parameters

λ	Weighting factor in the F-measure
IoU	Intersection over Union

Variables

$Loss_{box}$	Bounding box regression loss
$Loss_{obj}$	Objectness prediction loss
$Loss_{cls}$	Classification loss
TP	True positive
TN	True negative
FP	False positive
FN	False negative
P	Precision
$P - R$	Precision-recall
R	Recall
F	F-measure
F_1	F1-score
AP	Average precision
mAP	Mean average precision

Contents

List of Figures

List of Tables

1

Introduction

This chapter defines the background of the project and provides a problem description. Additionally, the purpose and aim, research questions, and the scope and delimitation of the project are presented in this chapter.

1.1 Background

Reliable material handling and quality assurance are essential in modern manufacturing and logistic processes, where even small errors can result in delays and large costs. As production systems become more complex and companies are expected to deliver faster and with higher precision, there is an increasing need for methods that improve process reliability.

Despite this, manufacturing today generally consists of both automated and manual elements [6]. With many industrial processes still handled manually and therefore susceptible to human error. In situations where operators must verify quantities and document packed material, mistakes can be costly and difficult to trace afterwards. This creates a need for solutions that can support or automate verification while also providing reliable documentation.

One technical area that has gained attention in this context is vision. Although machine vision has been used in industrial applications for more than 40 years [7], recent improvements in artificial intelligence (AI) have increased the number of ways it can be used [8]. Today, computer vision is not only used for inspection and quality control, but also for identification, tracking, and decision support in industrial settings [9; 10; 11]. For this reason, computer vision is of interest for industrial part counting. A vision-based system could potentially improve counting accuracy, reduce manual workload, and create traceable proof of packed material.

These developments make computer vision relevant in industrial processes where manual verification is still required and where mistakes are costly. One such case is the handling and verification of packed material in Volvo Trucks' complete knock down (CKD) operations. The Volvo Trucks plant in Tuve, Gothenburg, produces heavy trucks of the models FH16, FH, FM, and FMX. The plant also supplies parts and components to smaller assembly plants located in other parts of the world through CKD operations. In simpler terms, this means that the vehicle material is picked and packed, and thereafter shipped. If an error occurs, such as packing the wrong quantity of parts so that a box or container must be unpacked and corrected, the consequences can be significant. This affects not only the CKD department but also the Tuve plant as a whole, since the process is both costly and time-consuming. The impact is even greater if the container has already been shipped, as this results in delayed material to the customer and may cause delivery delays for the end customer. It is therefore essential to ensure that each shipment contains the correct quantity and the correct parts.

In today's production process, each shipment is manually kitted and controlled to ensure the correct quantity. The quantity control is a critical task, as errors can lead to customer complaints and extra costs. If the customers make claims and issue an unpacking control report (UCR), there is a process to determine who is responsible and liable for the deviation. This process begins with a root cause analysis (RCA) which not only aims to determine liability but also to provide necessary information to decide on actions to be taken. The burden of proof lies with the sending plant, in this case the Tuve Plant, and it is therefore important that there is available proof to contradict the claims if no wrongs have been done at the sending plant. Or, on the other hand, if the Tuve plant cannot show proof of no wrongdoing or if it becomes clear that errors have occurred at the plant, then Tuve must solve the problem. There are a few ways the problem can be solved, such as temporarily replacing the missing part(s) from stock at the receiving plant. However, worst case scenario is when it comes to the smaller receiving plants where there might not be any spare parts available and no new shipments on the way that they could borrow parts from. In this case, Tuve has to pack a new order with the missing part(s). Thereafter send the missing part(s), most often by plane, to get it there as soon as possible. An overview of this process is shown in Figure 1.1. Blue color indicates that the process is owned by the Tuve plant, purple indicates a transportation, and green is connected to the receiving plant.

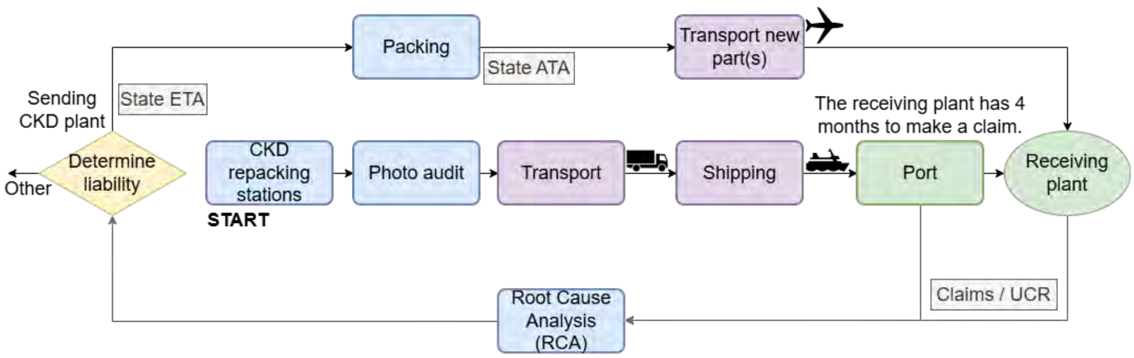


Figure 1.1: Conceptual model of the UCR process.

At Tuve, there are four identical CKD repacking stations staffed by several operators. While the repacking work is the same at each station, the operators rotate through other tasks, such as driving forklifts. See Figure 1.2 for a simple conceptual model of the area.

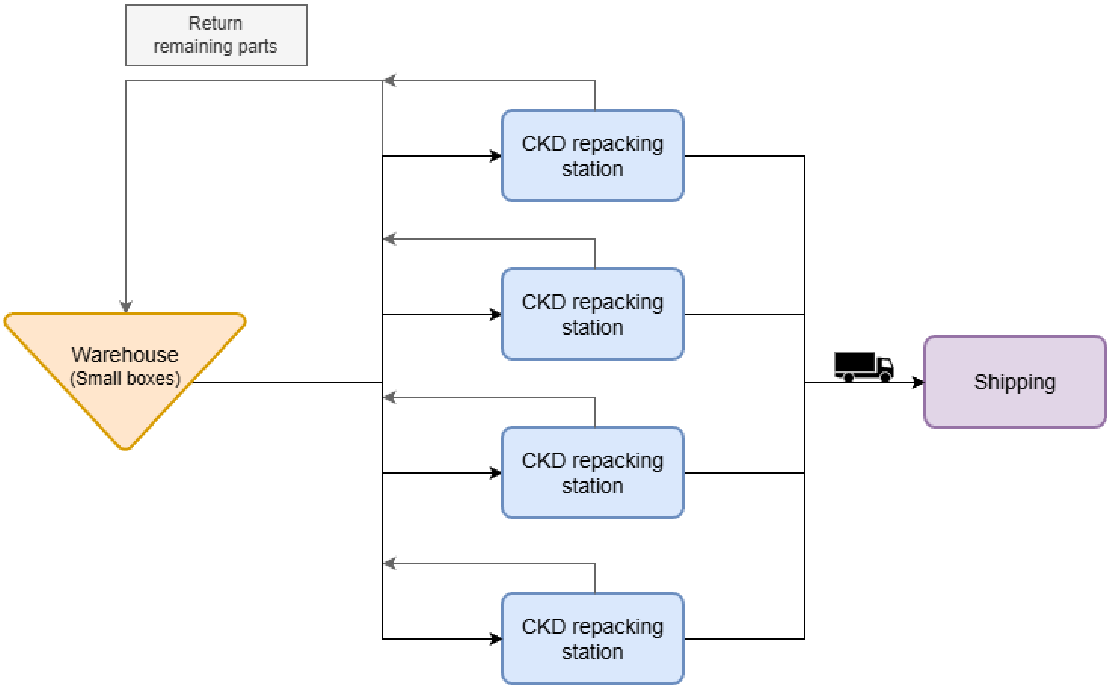


Figure 1.2: Conceptual model of the CKD process.

A department is responsible for the stations. An operator at such station is responsible for counting the articles and taking photographs that show the quantity of articles, their article numbers, and the box in which they are placed. Figure 1.3 illustrates how the operator might arrange the parts to make it easy to count them. Additionally, the pictures show that blue plastic lids are used as both work surface and background. The lids are the only designated area for taking pictures as of now, and there is no standardized method for arranging the parts. Instead, it is up to the operator to ensure that the pictures are of sufficient quality for later use.

Figure 1.3: Examples of pictures taken at the CKD stations.

In addition to this, the department is also responsible for a quality process in which all pictures are reviewed. In this step, a quality inspector sorts the pictures, renames them with the correct article numbers, and counts the material. Although this process aims to reduce errors, it is redundant, time-consuming, and allows for human error.

This process has remained pretty much unchanged for many years and historically there have only been done limited efforts to improve efficiency or introduce automation. The main reason is the investment cost of new equipment and process integration but also time constraints related to evaluation, testing, and implementation. However, maintaining the current manual process is hard to justify because it is inefficient, and requires an additional operator to review and sort the pictures. As production volumes increase and more material needs to be packed, the work process must be improved and streamlined in order to achieve a faster process and ensure reliable delivery.

1.2 Problem description

The current quantity verification process in the CKD repacking operation at Volvo Trucks Tuve is manual and depends on operator performance. This makes the process time-consuming, redundant, and vulnerable to human error.

In addition, the documentation is not standardized as the operator decides how the parts are arranged and photographed. As a result, the quality of the images and their usefulness as proof may vary. Incorrect quantity verification can lead to customer claims, extra administrative work, urgent deliveries, and delayed shipments. The current process is therefore inefficient, costly, and insufficient for increasing production volumes.

To address these challenges, Volvo Vision System (VVS) is evaluated as a potential solution for verifying delivery quantities and storing images as proof, although its performance in this context is unknown. Using a system from external suppliers is a potential solution, however it comes with some disadvantages and the main one is the high cost. Additionally, given the large number of different products that the system will have to handle, flexibility and adaptability are of great importance.

1.3 Purpose and aim

The purpose of this thesis is to better understand the potential and limitations of computer vision-based part counting at the CKD stations.

The aim of this thesis is to evaluate computer vision-based part counting at the CKD stations by reviewing the state of the art and assessing the performance of VVS at a CKD packing station.

1.4 Research questions

To achieve the aim of evaluating computer vision-based part counting at CKD stations, the study was structured around two research questions.

The first research question is motivated by the need to establish a theoretical foundation. Understanding current possibilities and limitations of existing approaches in industrial part counting is essential to evaluate potential solutions within the state of the art.

RQ1. What is the current state of the art in computer vision systems for industrial part counting?

The second research question builds on this by focusing on how a specific system (VVS) performs in practice.

RQ2. What is the performance of VVS in industrial part counting?

1.5 Scope and delimitation

The scope is limited to VVS as an industrial case. While the current state of the art in vision systems is reviewed, alternative solutions from external suppliers are not evaluated in detail, and they are not compared systematically to VVS.

The study focuses on CKD picking and packing operations in the small box warehouse. The analysis and proposed concept are developed for a single CKD packing station, which serves as an example for investigating integration into existing production environment. Other types of stations, processes, or plants are not studied.

The work is constrained by a project duration of 20 weeks and a hardware budget of 10 000 SEK. Within these constraints, VVS is not further developed or modified. Instead, the thesis is limited to evaluating its capabilities against identified requirements and proposing a station concept.

Finally, the thesis does not include long-term performance testing, so potential degradation of VVS accuracy or robustness over time is not addressed within this study.

2

Theory

This chapter presents the theoretical background relevant to the thesis. It introduces computer vision in industrial settings, different methods for object counting, the detection model You Only Look Once (YOLO), and Volvo's own computer vision platform VVS. The chapter also describes supporting technologies, including Azure, and relevant hardware.

2.1 Computer vision in manufacturing industry

Machine vision is not a new concept, it has been used in industrial applications for many years [7]. Already in the early 1980s, machine vision, despite its complexity and cost, began to be used in settings where 100% inspection was required since manual inspections were too slow and prone to error [7]. In the early 1990s, machine vision evolved from being very complex and expensive to becoming more affordable and thus more accessible [7]. Early applications of machine vision involved combining a camera with a computer to gather information about an object [7]. Furthermore, vision systems were used instead of manual labor for tasks such as determining the location of objects, inspecting dimensions, checking for the presence or absence of features, performing identification, and controlling defects. To meet the needs of new markets demanding lower costs, suppliers of vision systems had to develop their products during the 1990s to allow for higher throughput and/or lower price while still maintaining accuracy [7].

In the 2000s, the fourth industrial revolution began to shape industry, aiming to integrate advanced technologies that work with and for humans [12]. It became increasingly important to combine computers and automation to discover new possibilities for creating more efficient and reliable flows. Internet of things (IoT) is a cornerstone of Industry 4.0, allowing everyday objects to send and receive data, which enables them to cooperate not only with each other but also with humans [12].

Human vision is a vital aspect to consider in a production environment and a lot of tasks are depending on it [13]. However, there are limitations when it comes to what the humans are able to do. These limitations are often connected to cognitive ergonomic aspects such as fatigue or boredom. This is where computer vision can potentially aid the humans.

Machine vision developed into computer vision, which is a field of AI that enables machines to collect, interpret, and process visual information from the world to provide an output, typically using cameras and sensors. It allows real-time, non-contact inspections, and data-driven decision-making [14]. Today, computer vision development is an ongoing trend where the research about the topic cover areas such as computer theory, algorithm development, and realization [8]. In industrial settings where automation have been implemented, the image analysis, text recognition and industrial detection functions have been used to improve efficiency [8].

2.1.1 Applications of computer vision and common methods

Computer vision is used across many fields where automated analysis of visual data is needed. These include, for example, manufacturing, logistics, healthcare, security, and retail, where computer vision is used to support different types of visual inspection and decision-making tasks [15]. In these areas, computer vision is commonly applied to problems such as object recognition, scene understanding, defect detection, and medical image analysis [15].

To solve these tasks, both traditional machine learning (ML) methods and deep learning (DL) methods are used. Traditional methods include Support Vector Machines (SVM), decision trees, and K-Nearest Neighbors (KNN) [15]. More advanced approaches use DL models such as convolutional neural networks (CNNs) [15].

2.1.2 Hardware selection for computer vision systems

Selecting the right hardware for the task is essential when it comes to computer vision in a production setting. Aspects such as cost, environment, and speed are important to keep in mind.

The compatibility between lens and camera is determined by the lens mount which defines the mechanical interface between the lens and camera. Different types of mounts have different parameters such as flange focal distance and frame size [16].

How to mount the different components is also of great importance. Beyond aspects such as vibration, how each optical component is mounted in relation to each other is also essential [17]. They should not be able to move relative to each other. Failing to create a stable mounting for the hardware can directly lead to lowered image quality.

Industrial camera

There are two main types of cameras used for computer vision tasks: color cameras and monochrome cameras [16]. Color cameras are useful when it is important to distinguish between different colors in the image. They measure light intensity in three spectral bands, often referred to as the red, green, and blue (RGB) channels. Monochrome cameras, on the other hand, provide grayscale images. According to Volvo's internal training material, monochrome cameras can often provide higher image quality and more detail. Additionally, a colored image is not as easy to process since it has three channels (RGB) instead of one.

The shutter defines the way energy on pixels are captured. In other words, the shutter determines when the image sensor will be exposed to the light [18]. The two main options are global shutter and rolling shutter. A rolling shutter exposes the image pixels top to bottom which can make the image skewed if the parts are moving [16]. However, a rolling shutter is usually less expensive and is sufficient for applications like this one, where the parts are stationary when the image is taken.

The sensor is one of the most important camera components as its performance is strongly affected by it [18]. Important sensor properties include sensor size, resolution, and pixel size. Resolution refers to the number of active pixels in the sensor and better resolution leads to higher quality images [19]. A complementary metal oxide semiconductor (CMOS) sensor is the most common type of sensor and is also suitable for this project because of its low cost [16].

The interface defines the physical connection between the camera and the PC and affects how fast the data will be transferred [20]. The choice of interface therefore depends on the desired cable distance and, of course, the needed data rate to transfer the images.

The camera used for this project is area scan camera *AV ALVIUM 1800U-500C-CH-C* [4]. These types of cameras capture images in both the X and Y direction. Such cameras are suitable for processes in which the object does not move a lot which makes it ideal for this project. The sensor is a CMOS, which has benefits such as low cost and high speed [21]. Furthermore, the USB 3.0 interface provides enough transfer speed and still allows for shorter cable length and uncomplicated setup [20]. Lastly, a C-mount was suitable since it is widely used at Volvo, facilitating compatibility with available hardware [16]. See table 2.1 for some key camera properties.

Table 2.1: AV ALVIUM 1800U-500C-CH-C properties [4].

Properties	**Value**
Active sensor size (horizontal) [mm]	5.7
Active sensor size (vertical) [mm]	4.28
Connector a (video), b (power)	USB-3.0-Micro-B
Frame rate [Hz]	68
Lens mount	C-Mount
Pixel size [μm]	2.2
Resolution [MP]	5
Sensor format	1/2.5"
Sensor size diagonal [mm]	7.13
Sensor type	CMOS
Shutter type	Rolling shutter
Wavelength range	Color, visible

Lens

Choosing the lens for a project like this, with a large variety of parts, can be difficult. This is because the field of view (FOV) must be large enough to capture the biggest parts, but not so large that the smallest parts become difficult to capture clearly [22]. The working distance, i.e. the distance between the object and the camera, is also essential to know. Lastly, the focal length is the distance between the lens and the point where light converges, i.e. the focal point. In other words, this is how big the object is going to be in the picture.

The lens which was ordered for this project was the *HF12XA-5M* which is a high-resolution lens designed for "4DHR" [5] and key properties of the lens are summarized in Table 2.2. The specified working distance in Table 2.2, i.e. $\infty - 100$, is the lens focus capacity, meaning that it works well from 100 mm up to infinity. With the required FOV in this project, the working distance becomes 186 cm. It is made to be user-friendly as it take pictures with high resolution in many setups and conditions [5]. It is said to produce image sharpness consistently throughout the whole picture, avoiding degradation.

Additionally, the lens is made of a glass with extra-low dispersion characteristics that is said to be resistant to changes in the aperture value, i.e. how much light enters the camera [5]. This helps to maintain the high resolution. Moreover, it should be resistant to vibration and have passed both vibration and shock tests. This is of great importance to a project like this since there are a lot of movement at the stations and not a clear plan form the beginning where the camera will be placed.

Table 2.2: HF12XA-5M properties [5].

Properties	**Value**
Focal length [mm]	12
Aperture / Iris range [F. no]	F1.6-F16
Mount	C-mount
Working distance [mm]	$\infty - 100$
Angle of view(2/3")	40.1°×30.3°
Sensor size (std.)	2/3" (3.45 μm)
Sensor size (max.)	1/1.2" (4.5 μm)
TV distortion [%]	-1.26

Light

Lighting is an important aspect of computer vision systems. Finding optimal lighting is vital as features may be enhanced through one lighting method whilst others might become less distinguishable [22]. The direction and placement of the light source also affect how shadows will form, which may complicate the object detection process.

Designing a reliable vision system should therefore include designing specific lighting for the task rather than solely relying on ambient light [23]. This becomes particularly important when inspecting objects with varying materials that, for instance, reflects light more easily or is transparent [13]. In this project this can become challenging because of the high number of different parts and variation in texture, color, material, and geometry.

Moreover, stable and consistent lighting conditions are also important for achieving a reliable vision system and image acquisition [13]. This is because variations in ambient light, such as changes in the weather, may affect image quality and system robustness.

2.2 Vision-based object counting and its applications

Vision-based object counting can be performed using several different methods depending on the application and the complexity of the environment. The choice of method affects both the performance of the counting system and how well it can handle challenges. The following sections describe common counting methods, vision-based applications, and challenges related to large-scale counting systems.

2.2.1 Counting methods

Some of the methods that could be used to handle industrial part counting are: counting using object detection, segmentation based counting, density map based counting, few shot or exemplar based counting.

Object detection based counting

Object detection is used to identify, classify, and localize objects in pictures or videos [24]. The method identifies and locates each object with bounding boxes, and the objects are then counted based on the detections. Modern object detectors are typically based on DL and CNNs, such as two-stage detectors (e.g. Region-based CNN) or one-stage detectors (e.g. YOLO) that directly predict object locations and classes [24].

A benefit of object detection is that it not only counts the objects, but also determine where each object is located. This makes the approach suitable for applications where object information is required. Detection based counting tends to perform worse in situations where objects are occluded, very small, or packed, since it becomes harder to distinguish individual objects clearly.

Segmentation based counting

Segmentation refers to separating objects from the background by partitioning an image into regions whose pixels are similar with respect to features such as brightness, color, or texture [25]. However, this does not fully reflect the complexity of segmentation in real applications. A major difficulty is that the same object can have very different pixel values depending on illumination [25]. For example, if a sphere is lit from one side, one part of it will look brighter than the other. If the segmentation mainly depends on image intensity, this difference can make the object appear as several regions and it may then be counted as more than one object [25].

According to Davies [25], segmentation in industrial scenarios can be made more robust by designing the vision setup to control lighting and the surrounding environment. When the background illumination is uniform and the objects are separated from a contrasting background, segmentation can be performed using thresholding. However, even in controlled conditions, a practical challenge is to determine a threshold value that works well.

In more recent work, segmentation maps are described as providing pixel-level semantic labels and object boundaries, which can help models better understand image content [26]. This can improve segmentation in difficult cases, such as when objects overlap. It is further described that in autonomous driving, segmentation can help separate vehicles at different distances and support obstacle detection. Segmentation maps can also be useful for related tasks such as keypoint detection, object tracking, and scene understanding.

Density map based counting

Density map based counting is designed for scenarios with high object density, where detecting individual instances is not possible. Instead of identifying each object on their own, the model predicts a pixel density map over the image [9]. Then, the total object count is obtained by integrating the density map over the entire image.

This approach is particularly effective in crowded scenes with occlusion, as it in those scenarios is less time consuming than bounding-box detectors and does not rely on exact localization of individual objects [9]. Despite its robustness in crowded scenes, density map based counting it doesn't show exactly where objects are or where their edges are, so it's less useful for tasks that need detailed object information.

Few-shot or exemplar-based counting

Few-shot, or exemplar-based, counting methods estimate the number of objects in an image based on a small set of provided examples. Instead of learning to detect a fixed set of categories, the model is given one or more examples, such as images, and identifies similar objects in the scene [27]. Then, counting is performed either by localizing matching regions or by generating a density map guided by the exemplars.

An advantage of this approach is its flexibility. Since the model does not depend on predefined classes, it can generalize to previously unseen object types with no need for time consuming data collection and training. This makes it useful in applications where collecting large labeled datasets is not practical or where the target objects changes between tasks.

However, performance depends on how well the examples represent the objects in the scene [27]. Variations in scale, appearance, or occlusion can reduce accuracy. As a result, few-shot counting works best when the examples are representative and the objects in the target class do not vary much in appearance.

Category specific vs. class agnostic counting

Counting approaches can be divided into category-specific and class-agnostic methods, depending on whether they rely on predefined object classes. In category-specific counting, methods are formulated as detection, regression, classification, or localization tasks [28]. In practice, object detection and instance segmentation are typically category-specific, where each detected object is assigned a class label. These methods rely on that enough labeled data is available, and are usually limited to the categories seen during training. In contrast, class-agnostic counting methods estimate the number of objects using one or more reference examples instead of predefined classes [28].

A key difference between the two methods is that category-specific methods rely on features that separate classes, while class-agnostic methods use more general representations and stronger similarity measures to recognize objects that may look different [28].

2.2.2 Applications of vision-based object counting

Vision-based object counting is used in industrial environments where the number of objects is important for monitoring, control, or decision-making.

In manufacturing, object counting can be used to monitor production lines and verify the number of produced parts [10; 29]. This can support quality control and ensure that processes operate correctly. Counting in these scenes can be performed using different methods such as object detection or segmentation-based methods. In logistics, counting systems can be used to estimate stock levels, and support automated inventory monitoring [30; 11]. These environments often involve varying object positions and occlusions, which makes robust counting methods important for reliable predictions. In retail, object counting can be used to monitor shelf inventory and detect when products need to be restocked [31]. This helps improve efficiency in supply chain management and reduces the need of manual inspections.

Object counting systems are designed based on the complexity of the setting and the task, with simpler detection-based approaches used in structured environments and more advanced methods such as density-based or class-agnostic approaches used in cluttered or unstructured settings.

2.2.3 Challenges in large-scale object counting systems

In many real-world applications, counting systems need to handle a large amount of data. This creates challenges related to scalability, data handling, and data storage. Even though modern AI systems can be very advanced, their performance still depends on how well they can handle large amounts of data [32]. In practice, these systems need large amounts of training data, and it is often difficult to obtain relevant, high-quality datasets in industrial and manufacturing settings [33]. Therefore, system complexity becomes an important factor when applying computer vision in real-world environments, as it can create several challenges [34].

A key issue in this context is the increasing complexity of category-specific models as the number of object classes grows. Each additional object type requires new training data, and maintaining balanced and representative datasets becomes more difficult as more objects are added. In practice, object distributions are often long-tailed, meaning that a few categories have many samples while many others have only a few [35]. Furthermore, this imbalance can reduce performance, especially for rare or newly introduced object types. Similar challenges have been widely discussed in object detection systems with large and diverse object sets, where scalability and class imbalance become limitations [36; 37].

Another challenge is when new object types are introduced after deployment. In such cases, category-specific models typically require retraining for the new classes, which can be both costly and time-consuming [38]. This reduces the flexibility of the system and makes it less suitable for dynamic environments where the set of object categories is not fixed.

Class-agnostic approaches provide alternative strategies for addressing these challenges. By reducing or removing the dependency on predefined categories, these methods can more easily adapt to new object types without requiring full retraining of the model. However, this increased flexibility could have trade-offs in terms of accuracy or the level of detail in the output compared to category-specific approaches.

2.3 Volvo Vision System

All information about VVS has been gathered from Volvo's internal sources.

The VVS platform is based on YOLOv5 and is built using several connected modules that together support the complete process from model development and workflow creation to deployment in production. The main components are the VVS Portal, VVS Discovery, and VVS Edge. While the Model Development Platform and Azure are supporting tools. Each module has a specific role, but they are designed to work together, see Figure 2.1 for a simple overview.

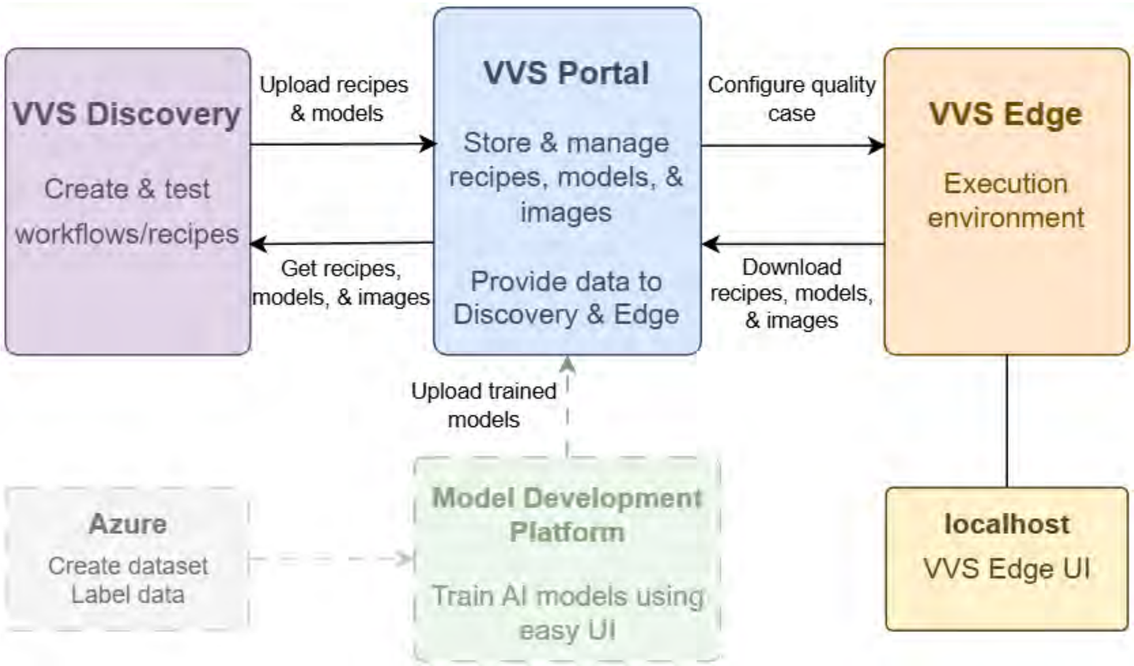


Figure 2.1: Overview of the VVS platform, its main components, and their relationships.

2.3.1 VVS Portal

The VVS Portal functions as the central hub of the platform. It is responsible for storing, managing, and distributing workflows (recipes), trained models, and images. It also acts as the communication bridge between VVS Discovery and VVS Edge. When workflows are created and validated in VVS Discovery, they are saved to the VVS Portal. Likewise, trained models developed in the Volvo Model Development Platform can be uploaded to the Portal for later use. The VVS Portal makes these resources available to both VVS Discovery and VVS Edge.

2.3.2 VVS Discovery

VVS Discovery is mainly focused on workflow creation. In this environment, users build recipes that define the sequence of operations required for inspection, analysis, or automation tasks. During development, VVS Discovery can access models and images stored in the Portal. This allows users to test workflows using available resources before deployment. Once a workflow has been validated and approved, it is saved back to the Portal.

2.3.3 Volvo Model Development Platform

Volvo Model Development Platform is a separate module and not a formal component of VVS, but rather a tool used to support users in training models. It is dedicated to AI model development and training, and its purpose is to simplify the training process by providing a simple user interface (UI). Users can train models using the platform and export the completed models. These trained models can then be uploaded to the Portal, where they become accessible for workflow testing in VVS Discovery.

2.3.4 VVS Edge

VVS Edge is the production environment where workflows are executed in real settings, such as factory stations, industrial PCs, or local laptops. Once the Portal has configured a station, VVS Edge downloads the workflow. The workflow can then be executed locally on the device.

VVS Edge is the final operational step of the platform, where developed solutions are transformed into real-world production processes. The UI for VVS Edge can be accessed via the localhost at port 5100. Through this interface, users can monitor, start, and manage workflows running on the local device.

2.3.5 YOLO and YOLOv5

YOLO is a family of real-time object detection models that detect objects in a single step [39]. Instead of first proposing regions and then classifying them, YOLO directly predicts bounding boxes and class probabilities from the full image [39]. This makes it fast while still providing good accuracy.

The input image is resized and divided into a grid, where each cell is responsible for detecting objects whose center lies inside it [40]. A CNN extracts image features, and the model predicts bounding boxes, confidence scores, and class probabilities for each cell [40]. Each bounding box is described by its center coordinates, width, height, and a confidence score [40]. Non-maximum suppression (NMS) is then used to remove overlapping detections and keep the most confident ones.

YOLOv5, introduced in 2020, is a later version of YOLO that keeps the same main idea but improves speed and performance [40]. Unlike YOLOv4, which is based on

Darknet, YOLOv5 is implemented in PyTorch, making it more flexible and easier to use [40; 39]. YOLOv5 is available in different model sizes, including YOLOv5 small and YOLOv5 medium, where the small one is faster while medium has a more balanced speed and accuracy.

Evaluation metrics: Loss function

YOLOv5 evaluates prediction quality using a loss function composed of three main components: bounding box regression loss, confidence prediction loss, and classification loss [40]. These components measure how closely the model's predictions match the ground truth. Therefore, they play an important role when assessing the overall model performance.

The bounding box regression loss, denoted as Loss_{box}, is calculated using Generalized Intersection over Union loss (GIoU_loss) [40]. This term measures how accurately the predicted bounding boxes align with the ground truth boxes. A lower Loss_{box} value indicates more precise localization.

The confidence prediction loss, denoted as Loss_{obj}, is calculated using *BCELogits_loss* [40]. This component evaluates how accurately the model predicts whether an object is present in a given bounding box. A lower Loss_{obj} value indicates more reliable object prediction.

The classification loss, denoted as Loss_{cls}, is calculated using BCE_loss [40]. This term measures how accurately the model predicts the correct object class. A lower Loss_{cls} value indicates better classification performance.

Together, these three loss components form the complete YOLOv5 loss function and provide a measure of localization accuracy, object presence confidence, and class prediction performance [40].

Evaluation metrics: Performance

When evaluating the performance of a computer vision method, a confusion matrix is commonly used to analyze prediction outcomes. It provides a structured representation of how predicted results compare to ground truth labels. In the context of object detection, predictions can be categorized into four outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [40]. See Figure 2.2 for a visual representation.

A true positive occurs when an object is correctly detected and assigned the correct label. A true negative represents the correct identification of the absence of an object in a region where no object is present. A false positive occurs when the model incorrectly detects an object where none exists, while a false negative corresponds to a missed detection, where an object is present but not identified by the model.

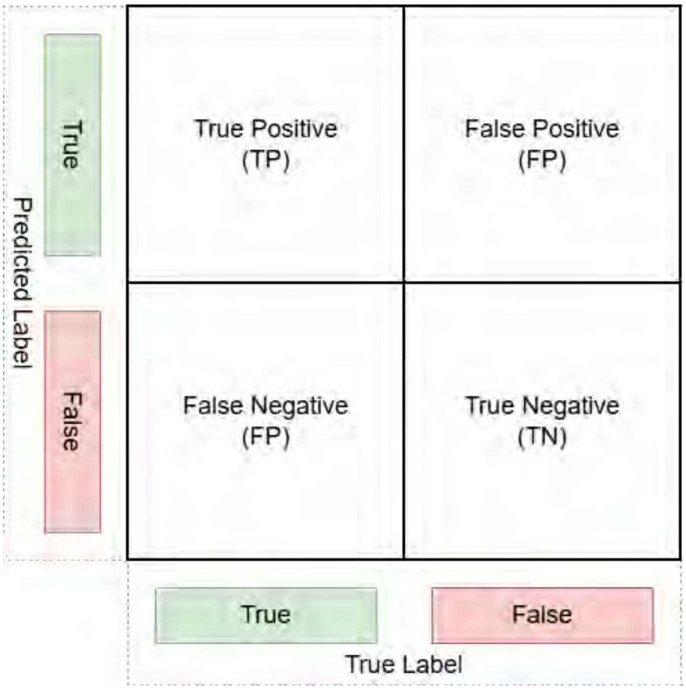


Figure 2.2: General structure of a confusion matrix.

Object detection models are commonly evaluated using standard performance metrics such as precision (P), recall (R), the precision-recall (P-R) curve, F-measure (F), average precision (AP), and mean average precision (mAP) [40]. Together, these metrics provide information about model accuracy.

Precision measures the proportion of correctly predicted positive detections among all positive predictions [40], and is defined as:

$$P = \frac{TP}{TP + FP} \tag{2.1}$$

Recall measures the proportion of actual positive instances that are correctly detected [40], and is defined as:

$$R = \frac{TP}{TP + FN} \tag{2.2}$$

While precision shows how reliable the detections are, recall indicates how completely the model identifies all relevant objects [40]. These two metrics are often in tension, meaning that improving one may lead to a decrease in the other [40].

To balance precision and recall, the F-measure is used, which represents their harmonic mean [40].

$$F = \frac{1}{\frac{\lambda}{P} + \frac{1-\lambda}{R}} \tag{2.3}$$

When $\lambda = 0.5$, the equation can be simplified to [40]:

$$F_1 = \frac{2PR}{P + R} \tag{2.4}$$

The relationship between precision and recall can also be visualized using the P–R curve, which illustrates model performance across different decision thresholds [40]. A model that achieves both high precision and high recall will produce a curve closer to the upper right corner, indicating better overall performance [40].

For a more comprehensive evaluation, Average Precision (AP) is computed as the area under the P–R curve for a single class [40]. Mean Average Precision (mAP) extends this concept by averaging AP over multiple classes, providing a single metric that summarizes overall detection performance [40]. Higher mAP values indicate better model accuracy [40]. In many applications, mAP is reported at a specific Intersection over Union (IoU) threshold, such as mAP@0.5 [40].

2.3.6 Azure

Microsoft Azure is a cloud computing platform developed by Microsoft and launched in 2010 [41]. Azure provides a wide range of services, including data storage, analytics, AI and ML development tools [41].

Azure Machine Learning is a part of Microsoft Azure, it is a cloud-based platform used for developing, training, deploying, and managing ML models [42]. It provides tools that support the different stages of the ML workflow, from data preparation to model deployment. The platform supports both models created directly in Azure Machine Learning and models developed with common open-source frameworks such as PyTorch, TensorFlow, and scikit-learn [42].

In this project, Azure Machine Learning was relevant as a cloud environment for handling image datasets and image labeling.

3 Methodology

This chapter describes the methodology used to investigate, evaluate, and develop a concept for verifying the number of parts in CKD shipments using VVS.

3.1 Design science research methodology

The study followed a design science research methodology (DSRM) approach, which is a structured way of designing and evaluating artifacts as solutions to real-world problems [2]. In this case, the artifact consists a of proposed station concept aimed at improving the quantity verification process at the CKD stations. Furthermore, DSRM was created for research projects to follow [2]. Thus, as this project does not aim at only developing a practical solution but instead evaluate and compare different alternative concepts based on research, DSRM was considered suitable for this study.

Moreover, Peffers et al. [2] specifically argues for following the DSRM in information systems. That is a system consisting of different components used to improve or support decisions and operations at the company [43], for instance, the proposed VVS solution such as in this case. This further motivates the choice of DSRM as an appropriate methodological approach.

Design science research (DSR) is guided by the following seven principles by Hevner et al. [44]:

1. The project should result in a concrete outcome such as a model, method, or system.
2. The objective should be to solve a real and relevant business problem by using technology.
3. The proposed solution must be carefully evaluated.
4. The research should contribute with new knowledge and insights.
5. The research should be carried out using a well-structured and systematic approach.
6. The solution should be developed iteratively and different alternatives should be explored and compared.

7. The results should be communicated effectively to both technical and non-technical audiences.

The DSRM process is based on six different steps which are described and illustrated in Figure 3.1 and based on the framework proposed and described by Peffers et al. [2]. This process was followed throughout the thesis work and each activity is described in the following sections.

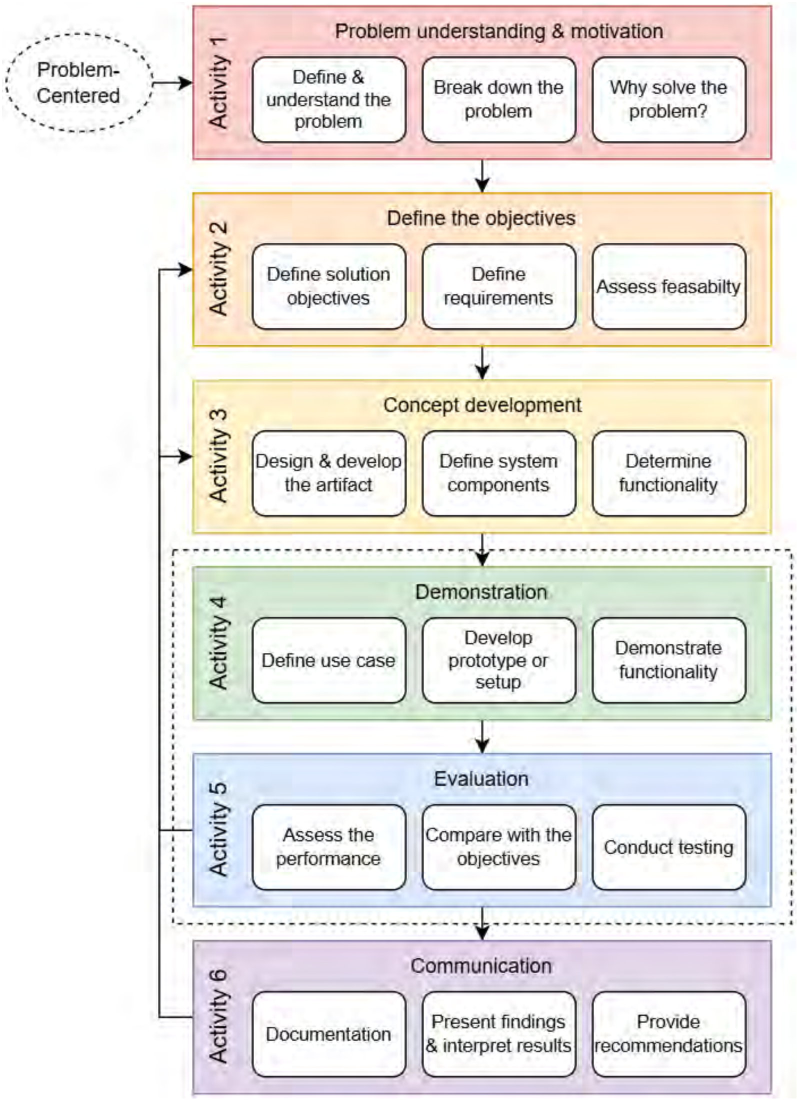


Figure 3.1: DSRM process model, adapted from the model created by Peffers et al. [2]

Following the DSRM ensures that the proposed solutions were developed and evaluated systematically. Although the activities are placed in order from 'Activity 1' to 'Activity 6', the possibility still remains when following DSRM to jump freely back and forth between them or even start with other activities than the first [2].

This study followed what Peffers et al. [2] calls a *problem-centered approach*, where the research might be initiated because of an observation of the problem. In this case, the study emerged after inefficiencies and waste where identified at the CKD stations.

3.2 Problem understanding & motivation

As the initial step of the DSRM approach the problem was investigated and clearly defined through a qualitative study consisting of a Gemba study, interviews, and observations. Additionally, this phase focused on understanding and motivating why the problem must be solved and the value of a potential solution.

3.2.1 Current station study

A current station study was conducted at selected CKD repackaging stations to understand the existing work process and identify limitations that would affect a vision-based solution. This was done to gather understanding of the practical setting in which VVS would potentially operate and thereby being able to address RQ2.

The study was inspired by Gemba walks, or "go and see", in which the work is observed where it is performed in order to gain a direct understanding of the process [45]. This approach was chosen to study how parts are currently counted, photographed, reviewed, and documented in practice, and to identify challenges and improvement opportunities in the current workflow. Gemba walks are commonly used to better understand a process through direct observation [45]. Additionally, unstructured interviews with operators and team leaders were conducted during the Gemba walks. These types of interviews does not involve any pre-planned questions and can be compared to regular, unplanned conversations [46].

The following types of data were collected:

- **Process steps and responsibilities**, documented with process maps from kitting to quality review.
- **Operator interaction**, including how items are handled and placed, how pictures are taken, and where errors might occur.
- **Environmental constraints**, such as available space, mounting possibilities, and lighting conditions.
- **Variation drivers**, including article variation, packaging types, part geometries, and common counting challenges.
- **Quality requirements**, including what proof is required, how those images are currently stored, and what requirements a future solution must fulfill.

3.3 Defining the objectives

The next phase of the DSRM approach includes defining the solution objectives based on the identified problem [2]. In this study, this included identifying key stakeholders and requirements for a potential VVS-based station concept, as well as specifying desired improvements and constraints.

3.3.1 Stakeholder analysis

A stakeholder analysis was carried out to identify the actors affected by the project and to understand their interests, influence, and importance to the project. According to stakeholder theory, an organization should consider not only internal decision-makers, but also other parties who may be affected [47]. In this thesis, the stakeholder analysis was used to create a structured overview of the people connected to the CKD station process.

The analysis started by identifying both internal and external stakeholders. Internal stakeholders included people within the organization who are directly involved in the process, such as operators, quality inspectors, and team leaders [47]. External stakeholders included actors outside the internal process who could still be affected by the solution or have an interest in its implementation [47].

The stakeholder analysis in this thesis was mainly descriptive. This means that its purpose was to map stakeholder relationships and provide an understanding of how different groups may view, support, or oppose a project [47]. This approach was suitable because the aim was to identify relevant stakeholders and understand their roles in relation to a proposed solution.

After the stakeholders had been identified, their main interests were analyzed. This included considering how they could be affected by process changes and what expectations they might have.

3.3.2 Requirement identification

Requirements were collected through interviews and observations, and then structured into the following categories:

- **Station layout**, including space constraints, placement relative to the station, ergonomics, safety, and operator interaction.

- **Station components**, such as camera, optics, lighting, fixtures, computing infrastructure, and communication.

The interviews were semi-structured, meaning that the questions were prepared in advance but still allowed additional questions to be asked during the interviews based on the responses and discussion with the interviewees [46]. A semi-structured interview is suitable since the intention of the interviews is to gain insights into the interviewees' personal views [48]. The interview questions are provided in Appendix A.1.

3.3.3 Literature review

Based on the defined requirements, a literature study was conducted to understand the state of the art of computer vision in an industrial setting, i.e. to investigate RQ1. The findings were then used to guide and support method selection and station design choices.

It was conducted as a systematic literature review, which is a structured method for finding, assessing, and interpreting available research related to a research question [49]. Furthermore, a systematic review is used to summarize existing evidence and identify gaps. Additionally, it is important to provide research value. The review followed the methodology for planning and conducting a systematic literature review proposed by Kitchenham [49].

The findings from the literature review were used to refine the solution objectives but also to guide the next phase, i.e. concept development.

Planning the review

There are two stages when planning the review: identifying the need for a review and making a review protocol [49].

The first stage involved determining whether a new systematic literature review was necessary. This included examining the existing research to see if similar reviews have been conducted, assessing their scope, quality, and identifying any gaps or limitations [49]. If current reviews are outdated, too general, focus on a different context, or do not address the research question of interest, this provides a motivation for conducting a new review.

The second stage involved developing a review protocol to reduce researcher bias, following the guidelines proposed by [49]. The protocol defined the background, research question, search strategy, selection criteria, quality criteria, data extraction process, and synthesis method used in the literature review.

The review was motivated by the observation that vision-based systems for industrial part counting have reached a relatively high technical maturity, but are still not widely adopted in Volvo Tuve's production environment. Therefore, a systematic literature review was conducted to summarize the state of the art and investigate whether computer vision could be a suitable solution for industrial part counting. The review protocol was developed around RQ1.

To identify relevant literature, a search strategy was developed using the Scopus database, chosen for its ability to deliver peer-reviewed scientific publications including journal articles and conference papers. The search terms were formed around vision, object or part counting, and industrial or production environments. The complete search strategy is presented in Table 3.1.

Studies were included if they proposed, evaluated, or reviewed computer vision-based methods in industrial settings where computer vision was used for counting or verification purposes. Studies were excluded if they were not written in English, not related to industrial environments, not focused on counting or verification, or did not involve computer vision techniques. Only peer-reviewed publications were considered to ensure scientific quality.

Table 3.1: Scopus search strategy.

Search within	Article title, Abstract, Keywords
Search documents	("computer vision" OR "machine vision" OR vision*) AND (industr* OR manufactur* OR produc*) AND (("part" OR "object" OR "item") W/0 count*)
Filter by language	English
Filter by subject area	Cumputer science, Engineering, Decision science, Multidisiplinary, Business, management and accounting
Date of search	2026-03-05

Conducting the review

A systematic literature review should be conducted in five main stages [49]. First, relevant research was identified for this study based on the search strategy defined in the review protocol, with the aim of finding as many potentially relevant studies as possible. Second, the studies were screened using the inclusion and exclusion criteria to determine which studies would be included in the review. Third, the quality of the included studies was assessed according to the study quality criteria specified in the literature review protocol. Fourth, data relevant to the research question were extracted from each study using a structured data extraction form. For each selected study, data extraction focused on bibliographic information, context, and the main points of the study. Finally, the extracted data was analyzed using a qualitative approach.

3.4 Concept development

This phase of the DSRM approach regards the actual design and development of the proposed solution [2]. In this case a VVS-based station concept. This involved exploring and formulating alternative station concepts, as well as finally selecting and developing a solution based on the defined objectives and constraints.

3.4.1 Hardware and settings

As part of the third activity, i.e. the development stage, different experiments were conducted. These experiments were part of the design and development stage of the DSRM approach as the test results were used to develop the actual station concept [2] and to investigate RQ2.

These tests were conducted to identify feasible ranges for camera placement (e.g. working distance, angle, FOV, and stability), lighting conditions, and operator interaction. The latter refers to how parts are presented to the camera and how the new work process may affect operators. They were also used to support the development of the station concept and guide design decisions. Additionally, these initial experiments were conducted to gather a better understanding of VVS and its capabilities. This included connecting the camera to test its functionalities and navigating in the system.

So, to start testing, the camera was positioned on a table in front of a test object and images were captured while varying the distance and angles. Different properties, such as aperture, zoom, and acquisition control, were changed throughout the testing. The lighting conditions were also tested by using natural light from the windows to see how lighting variation, in this case changes due to weather conditions, would affect the camera. These tests were conducted to gather a deeper understanding of the camera configuration and to evaluate the sensitivity of the camera and VVS. The captured images were uploaded to the software to evaluate different detection abilities.

After a general understanding of VVS had been gathered, the next phase of experimenting was initiated: building a first model. For this, a test rig was built to hold the camera in a position that satisfied the working distance and FOV requirements. The setup included the camera, lens, a blue lid for background, and the preliminary camera rig. The complete setup in the lab is shown in Figure 3.2.

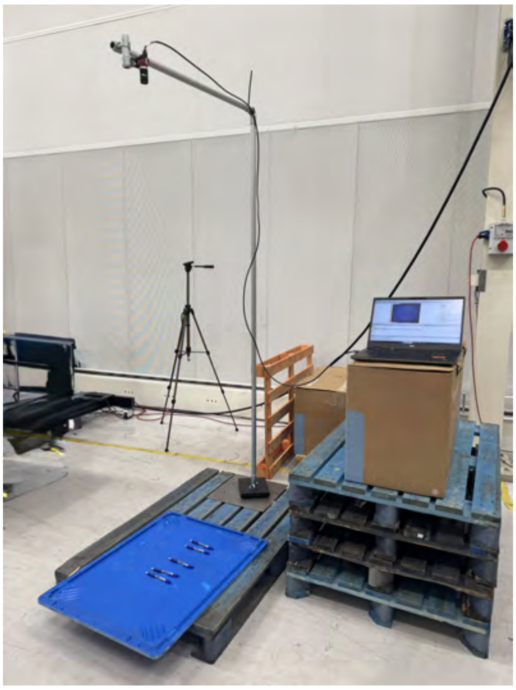


Figure 3.2: Test setup at the lab

This test was used to evaluate the camera under controlled conditions, while the daylight test near a window was performed separately without this setup.

Then, images were captured of some test objects, such as pens and stripes. These images were uploaded to Azure for labeling and then to the development platform to analyze the dataset, train the model, and lastly, to test the model. From this, data was collected and used to support both development and evaluation of the VVS-based counting solution.

3.4.2 Station concept development

The station concept was developed based on the defined requirements and test results. The objective was to design a solution that would include VVS, improve the current process, and allow for simple implementation at the stations. Especially important was to ensure that the solution would not lead to unnecessary complexity or added workload for the operators.

The concept includes the proposed station layout and operator interaction steps. This includes how parts would be placed, how the images would be captured, and potential operator feedback. Moreover, integration considerations including image storage for proof and overall traceability was another important aspect when developing the concept ideas.

During this phase, multiple solution ideas were also considered and evaluated based on the defined requirements. However, since this thesis focuses on assessing the feasibility of using VVS at the CKD stations, greater emphasis was placed on the VVS-based solution. The other ideas were investigated to enable comparison and to provide Volvo with the best possible final recommendation.

This activity also included evaluating the different solutions and the current one by identifying trade-offs based on requirements such as cost, accuracy, and robustness. For instance, one solution that could potentially be really accurate might instead be too costly or time-consuming.

3.5 Demonstration and evaluation

The demonstration phase of the DSRM aims to show how the station concept can be used to address the identified problem [2]. In this study, the usability of the developed station concept together with VVS was demonstrated through further testing, i.e. conducting the station-based tests.

At the same time, the main tests were also used to evaluate the proposed solution. The evaluation phase of the DSRM approach involves measuring and evaluating the created artifact [2], in this case the proposed station concept and VVS. Therefore, the test results were compared with the identified objectives and success criteria.

As the main tests were done to both demonstrate the artifact usability and to evaluate its performance at the same time, the demonstration and evaluation phases are described together in this report. This is also visualized by the dotted square in Figure 3.1.

3.5.1 Station-based testing

The main testing was conducted at one of the CKD stations to evaluate the solution under variation representative of everyday production. Examples of such variation includes different article numbers and packaging configurations, and repeatability tests where the same condition is tested multiple times. This was done to verify performance against the requirements and to contribute to answering RQ2. The same hardware was used as in the previous tests, except for a change of background. The setup is shown in Figure 3.3.

Figure 3.3: Test setup at the station.

A total of 11 different articles were used for these main tests and those were chosen depending on what parts were available at the stations during the time of the testing. A total of 100 images were collected of each of the 11 parts under normal packing conditions at one of the stations. A typical use case was defined where an operator places parts on the station surface and ensures that the parts are not overlapping or touching. After which the operator presses a button on the computer and the system captures the image. Several example scenarios were conducted to illustrate how the system operates under different conditions, including variations in part type and arrangement. The same camera settings and background were used for each image. The result was then stored with a consistent naming method, including the article number.

Thereafter, the images were uploaded to Azure that were used to label the data before uploading it to the Volvo Model Development Platform, which were then used to train and compare different models. Two different versions of models were developed, YOLOv5 small and YOLOv5 medium. Both of these were trained on 50, 70, and 90 images per class, which then resulted in six different trained models.

The system performance was then evaluated using the metrics explained in Section 2.3.5, including losses, confusion matrices, precision, recall, F_1-score, and precision-recall.

After that, the models were uploaded to the VVS Portal and could thereby be used by the VVS platform.

3.6 Communication

The last part of the DSRM process is the communication [44]. Throughout the project, documentation was updated and organized to simplify handover. This included Gemba notes and current-state process maps, requirement specification and stakeholder analysis, test plans, test logs, and result summaries, dataset description and data collection results. Furthermore, a station concept description, recommendations for future development, and, of course, the master thesis report.

4

Results

This chapter presents the main findings of the study, including the results from the current station study, literature review, hardware and settings experimentation, main tests, and the development of the station concepts.

4.1 Current station study

Here the findings from the Gemba inspired current station study are presented, including process maps over the current processes and operator tasks, factors affecting the current and future state.

4.1.1 Current state

Process steps and responsibilities

The process steps are illustrated in Figures 4.1 to 4.4. One operator is responsible for packing a new batch, as shown in Figure 4.1. A second operator is responsible for image storage and quantity control, as shown in Figure 4.2. If an error is detected or the image quality is insufficient, the procedures shown in Figures 4.3 and 4.4 are followed. In these cases, additional personnel may be involved beyond the operator performing the quantity control, such as team leaders and, potentially, other members of the logistics chain.

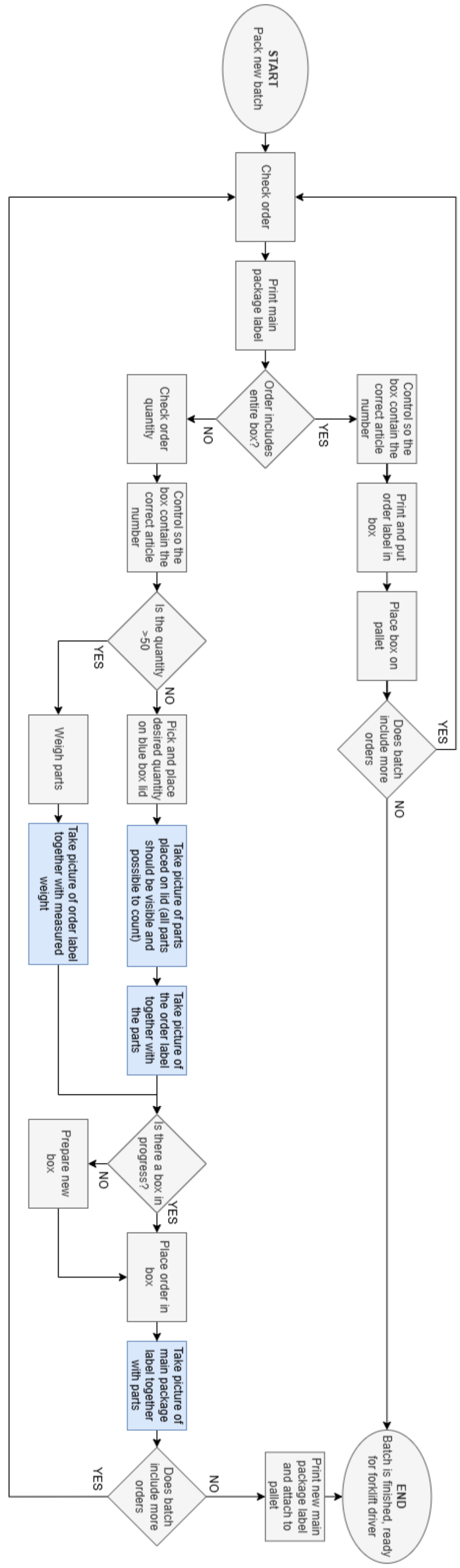


Figure 4.1: Process map: CKD picking & picture taking.

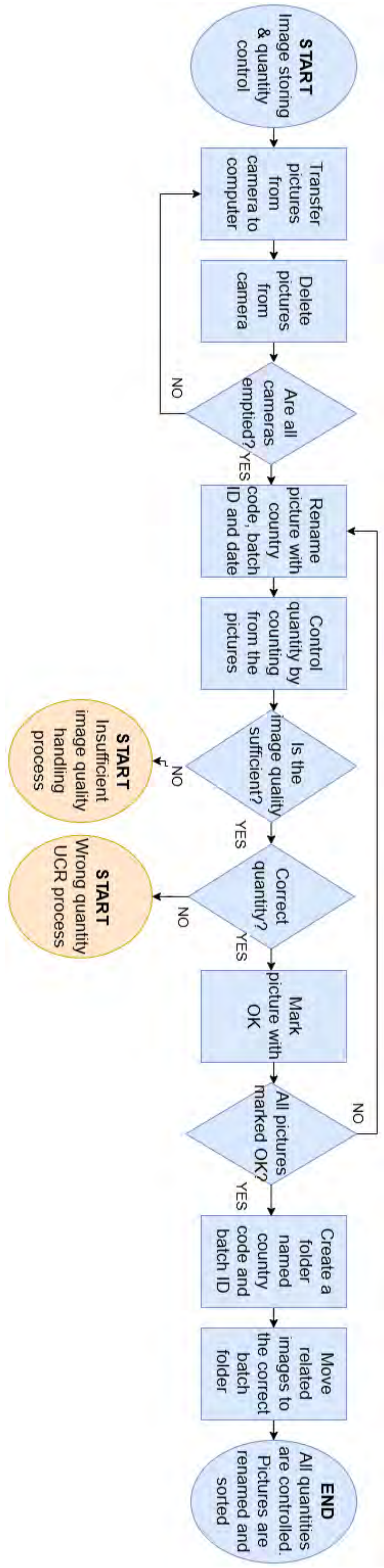


Figure 4.2: Process map: CKD image storing & quantity control.

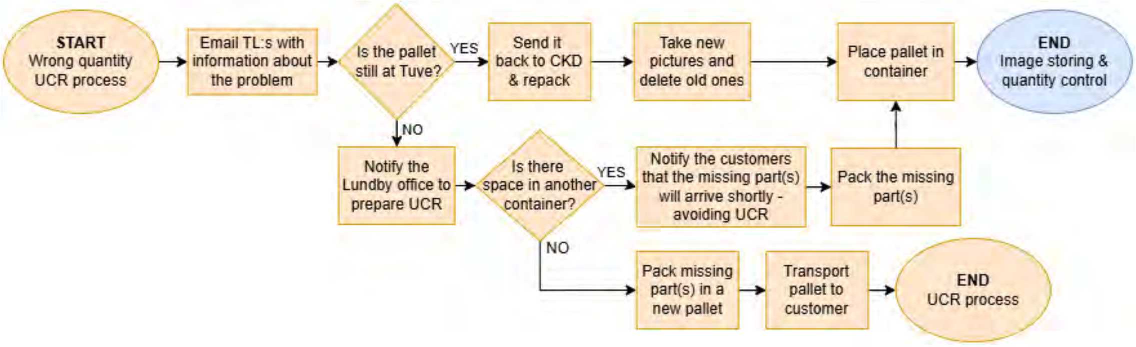


Figure 4.3: Process map: CKD wrong quantity UCR process.

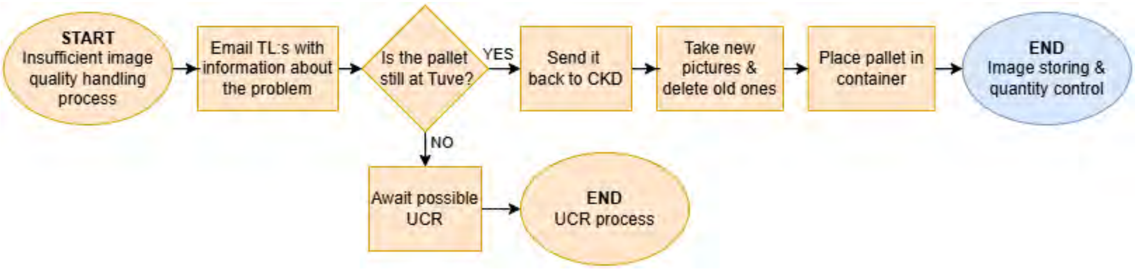


Figure 4.4: Process map: CKD insufficient image quality handling process.

Operator interaction

At present, there is no standardized method for how operators should place the parts when taking images. The only requirement is that they should "take a photo that clearly shows the material and the quantity, as well as our label with the stamp on it. Check the photo in the camera." For weighed material, the instruction is to "take a photo that clearly shows: the type of material, the quantity displayed on the scale, and our label with the stamp on it. Check the photo in the camera." After the material has been packed in a blue box, the instruction is to "photograph the flag and the material when the material is in the box. Check the photo in the camera."

Environmental constraints

There are four CKD stations, each equipped with operator workstations, conveyors, pallet handling spaces, and a traverse system spanning over all four stations. The basic layout of the stations cannot be changed. Therefore, any proposed solution must not obstruct either the traverse system or the operators.

The workstations are illuminated by industrial lighting mounted above the stations, at a height of several meters from the work area. As a result, the lighting is distributed from overhead rather than from sources positioned close to the operator or the parts being handled. Additionally, there is a window far above the stations, even though it is rather small it provides some natural lightning.

Variation drivers

There is a wide variation in the articles picked and packed at the CKD stations. Approximately 1300 articles are currently active in this repackaging process, differing in size, material, weight, and texture. The range includes everything from small screws and washers to long cables and large brackets. Figure 4.5 illustrates the large variation between two articles, which differ significantly in terms of material, size, and geometry. While some materials differ significantly from one another, others are very similar. Among the 1300 articles, there are several types of screws, hoses, and brackets, for example, and the differences between individual items within these categories may be small.

(a) Hose clamps **(b)** Pipes and hoses

Figure 4.5: Examples of article variation at CKD station.

The material arrives at the station in standardized blue boxes. However, the internal packaging may vary. As shown above, some materials are placed freely inside the box, while more sensitive materials may require additional packaging for protection against damage or corrosion. Figure 4.6 illustrates an example of material arriving at the station packed in plastic bags.

Figure 4.6: Example of material arriving to station packed in plastic bags.

Quality requirements

As previously shown in Figure 4.2, all images are renamed using the country code, batch ID, and date. They are then sorted into folders corresponding to the relevant country code and batch ID.

4.2 Stakeholder analysis

A stakeholder analysis was carried out to identify the main actors connected to the project and to understand their interests. The analysis included both internal and external stakeholders. The results are presented in Table 4.1.

Table 4.1: Results of the stakeholder analysis.

Stakeholder	**Project role**	**Main interests**
Volvo Tuve	Owns the process	Improve shipment traceability and accuracy, reduce customer complaints and costs, and standardize the kitting process
CKD: team leaders	Support daily operations, coordinate operators, and have contact with other departments	Stable workflow, reaching production goals, training material, clear instructions, and reduced errors
CKD: operators	Perform packing, picture handling, and quantity control in daily work	Simple and fast interaction and feedback, ergonomic work, clear instructions, and reduced errors

Maintenance	Supports maintenance and long-term functionality of technical equipment	Reliable equipment, clear maintenance instructions, and avoiding solutions that are difficult to maintain
Future employees	May use, learn, or work with the solution in the future	Clear documentation, clear instructions, easy use, and easy on boarding
VVS team	Contribute with technical solutions an knowledge	Clear requirements and an opportunity to demonstrate how VVS can be used
Customers	Receive the packed shipments	Improved shipment accuracy and easy access to proof of packing
Supervisors: Volvo	Guide the project from the company side and ensure relevance to Volvo	Thesis delivery that meets the project goals and contributes to the process
Supervisors: Chalmers	Guide the academic work and support the thesis process	Meeting academic requirements and ensuring that the work is presentable
Examiner	Evaluate if the thesis fulfills academic requirements	Requirements are met
The students	Carry out the study and deliver the thesis	Deliver a successful thesis, learn about VVS, build contacts, and gain project experience

4.3 Requirement identification

This chapter presents the identified requirements, more specifically, the findings from the interviews and observations.

A requirement is that the solution should not increase the cycle time, although reducing it would be beneficial. Furthermore, the station layout should remain mostly unchanged. It is acceptable to add structures for mounting cameras and other equipment, but major modifications should be avoided. The traverse system must remain operational. The solution must also comply with Volvo's safety and ergonomic requirements. Additionally, the operator should not be negatively affected by the new process. The solution should be easy to use, minimize the risk of errors, and not obstruct the operator's work.

No specific hardware components were determined to be a requirement. Several cameras, lenses, and setup alternatives could be suitable. The main requirement is that all components must be robust enough to operate in an industrial environment.

4.3.1 Interview findings

Interviewees:

- **Interviewee 1:** Senior Logistics Engineer
- **Interviewee 2:** UCR Quality Manager - CKD Operations

Tuve has the burden of proof and must ensure traceability in their process. Meaning that the Tuve factory must ensure that the correct articles and quantity is being sent and if a claim is made they must be able to prove that they sent everything in a correct manner. Otherwise Tuve must send new parts as soon as possible to the customer, which becomes very costly.

Before 2014, Tuve had no real knowledge or proof of what they were sending to their customers. The operators had all the responsibility and there were no system to ensure that they did everything right. That year, the department made the decision to start looking into their quality process by starting to re-count the number of articles at the stations two or even three times just to ensure correct quantity. Despite these extra checks, the amount of claims remained the same. This led to Tuve realizing that they needed to do something different.

Around 2018 they implemented internal audits where they did random checks of the packages after the CKD stations and before sending the orders. From this, they could see a clear improvement of the amount of claims but also as time went on, the amount of wrongly packaged quantity decreased, pointing towards the operators being extra careful at the CKD stations. Additionally, this saved money as the operators could fix the mistake before it ever reached the customer. It was also beneficial in the sense that the operators got much quicker feedback of any wrongdoings. Before, claims happened multiple months after the mistake had been made at the stations.

After a few years the audits positive affect started to decrease and Tuve could also see how some customers could make claims even when audits had been made, in other words Tuve knew that the correct article and quantity had been shipped but they could not prove it. This led to the need for another solution that prevented wrongly packed orders to be sent but also saved proof of quality. They then decided to implement the photo audits. With more knowledge and control of their own processes, Tuve could then put some pressure back on the customer factories to look into their own processes to try and find gaps there where claims get made when that should not happen. The interviewees pointed out that Tuve wants to avoid the unneccessary claims and want to help the customers in various ways to ensure a complete and stable process from the CKD stations all the way to finished trucks in the receiving countries.

CKD statistics

The interviewees emphasized the large number of shipments processed at the CKD stations each month, where each shipment contains all parts needed for a complete truck. This further highlights the importance of ensuring a process with minimal errors. Specific statistics are presented in Table 4.2.

Table 4.2: Statistics of CKD shipment volume in 2025.

Country	**Total volume**	**Average per month**
Country 1	424	35
Country 2	116	10
Country 3	176	15
Total	716	60

UCR statistics

Before the department started working more with quality in 2014, the average UCR claim per truck and month was 1.5-2. Today their target number is 0.17 which is a significant improvement and does not leave much room for errors.

There are a few different types of claims and they get divided into eight groups named code 1 up to code 8 and the cost of a customer claim varies a lot between the groups. In 2025, the CKD department at Tuve received 410 claims in total. There are eight different codes for possible claims but the two that are potentially most affected by this projects is Code 1 and Code 6. Code 1 is missing article which was the reason for 210 of those claims and was estimated to cost Tuve at least 600,000 SEK. Code 6 represents sending the wrong article, which was responsible for 73 of the claims and led to an estimated cost of at least 700,000 SEK. These costs are calculated when not taking into account the extra shipping or flight costs which is usually a lot.

Table 4.3: Tuve statistics of UCR claims in 2025.

Claim type	**Description**	**Number of claims**
Code 1	Missing article(s)	210
Code 3	Packing damage	104
Code 6	Wrong article	73
Code 7	Excess material	23

Cost for the photo audits

Having an operator costs 855,184 SEK/year. Apart from the more straightforward costs, it costs a lot to to spend the extra time to do the whole photo audit process. Having an operator do all the manual admin work instead of value adding work is a complete waste.

4.4 Literature review

The search string defined in Table 3.1 resulted in 124 documents. After applying inclusion and exclusion criteria, the remaining studies were screened through abstract and full-text review. The overall result of the selection process is summarized in the PRISMA diagram in Figure 4.7.

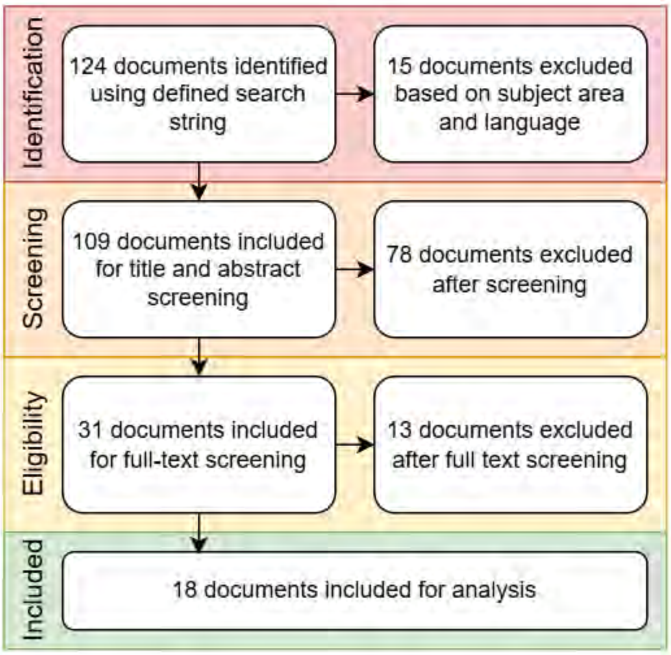


Figure 4.7: PRISMA diagram showing the literature selection process, adapted from [3].

Table 4.4 presents the studies identified through the literature review and shows the wide range of approaches used for object counting across different areas. The reviewed studies are published between 2015 and 2025, and show an evolution from more traditional methods toward DL based methods.

Table 4.4: Extracted data from selected literature.

Reference	Year	Context	Short description
[26]	2025	Open-world object counting	Introduces TrueCount, a multimodal framework that combines RGB images, segmentation, depth, and text/visual prompts with large vision language models (LVLM) to perform general, open-world object counting in complex scenes with stacked and occluded products.

[9]	2025	Small parts in aircraft manufacturing	Presents a density-map regression system to count small, visually similar components (e.g. screws, bolts, washers) in cluttered scenes. Shows that combining synthetic and real data improves performance when manual labeling is limited.
[50]	2025	Casting industry	Presents DeepMachining, a dual-branch YOLO-inspired architecture to detect and count overlapping castings, achieving about 97.8% counting accuracy but with high model complexity.
[30]	2025	Steel sheet stacks	Proposes TSNet, a two-stream network using RGB and gradient maps to emphasize contours for accurate counting of tightly packed steel sheets. Discusses detection-, regression-, and density-based methods and their trade-offs for stacked sheet counting.
[51]	2024	Agriculture	Proposes DMseg-Count for wheat spike counting, a density-map regression solution to better handle occlusion, overlap, and illumination issues. Compares multiple DL models including YOLOv8 and Faster R-CNN.
[31]	2024	Retail shelf monitoring	Develops a YOLOv5-based system for object identification, item counting, and product categorization in cluttered retail shelf images. Discusses overlapping objects, and compares three YOLOv5 variants.
[10]	2023	Shoe manufacturing	Uses YOLOv4-tiny with DeepSORT and OpenCV on CCTV video to detect, track, and count safety shoes in real time for work in progress monitoring, achieving about 92.6% mAP.
[52]	2023	Mobile phone object counting	Proposes a mobile phone based system for object detection and counting using Otsu thresholding, Hough transformations, edge detection and image-processing algorithms.

[29]	2023	Conveyor object counting	Proposes a real-time machine vision tool that detects, counts, and sorts identical plumbing objects on a conveyor into bins using a minimum-distance classifier. Argues this classifier is sufficient under controlled conditions.
[27]	2023	Few-shot object counting	Addresses few-shot counting where only a few support images are available, introducing similarity comparison and feature enhancement modules to count densely packed objects in query images.
[53]	2023	Human-in-the-loop counting	Introduces an interactive framework where a density-based counter is refined using user-provided regions and ranges. Improves performance of class-agnostic visual counters for a modest amount of human input.
[28]	2022	Class-agnostic counting	Proposes a similarity-aware framework for class-agnostic counting that jointly learns feature representations and similarity metrics end-to-end, improving extract-and-match pipelines and enabling counting of categories.
[54]	2021	Retail / produce checkout	Combines a scale and webcam with lightweight deep models (MobileNetV2/EfficientDet/EfficientNet+ Decision Tree) to classify and count fruits and vegetables, arguing that object counting without precise localization is sufficient and more efficient at checkout.
[55]	2020	Conveyor / rectangular items	Implements an OpenCV/C++ image-processing pipeline, including Hough transform, to segment, detect, and count partially occluded rectangular objects (demonstrated with playing cards on a dark background) on a conveyor-like setup.
[56]	2019	Vibratory parts feeder	Introduces a 1D intensity distribution based counting method for parts on a vibratory feeder, improving robustness when there are no gaps between parts and it avoids full 2D template matching assumptions.

[57]	2018	Multi-domain object counting	Proposes a patch-based CNN regressor with domain-specific parameters to adapt counting models across diverse visual domains (people, animals, cells) without forgetting, improving counting accuracy across domains by summing patch-wise counts.
[11]	2017	Inventory control	Presents a machine vision approach for real-time inventory monitoring using SURF-based appearance models as visual signatures to identify, localize, and count mainly rectangular objects in video streams.
[58]	2015	Object counting	Proposes object counting via colour-histogram-based segmentation and spatial filters, introducing fuzzy colour histograms for robustness under non-uniform illumination and using size and dip filters for segmentation and counting.

Earlier studies mainly used traditional computer vision methods [11; 55; 56; 58]. These include color histograms, Hough transforms, feature descriptors such as SURF, BLOB analysis, and intensity-based analysis. Many of these methods were developed for relatively controlled settings, for example with stable lighting, simple backgrounds, predefined object motion, or visible spacing between objects. They often emphasized simplicity and speed, but were generally less robust in cluttered scenes with occlusion or overlap.

Later, DL methods became more common in the reviewed literature. Studies such as [28], [27], and [53] focus more on class-agnostic, few-shot, or interactive counting. These systems are designed to improve generalization, and some can count unseen object types or work with only a small number of exemplars. They often use CNNs, similarity learning, and density-maps to handle crowded scenes and occlusions more effectively.

Many of the reviewed studies focus on industrial use cases, such as aircraft manufacturing [9], casting foundries [50], steel sheet counting [30], conveyor/part feeder systems [29; 55; 56], and production monitoring [10]. A common challenge in these areas is that objects often overlap, are partially hidden, densely packed, or visually similar. To address this, researchers use a range of methods, including YOLO-based object detectors, density map methods, two-stream networks, and combinations of feature extraction and tracking techniques.

The literature also shows a trend toward more advanced and flexible systems.

Some recent work uses large vision-language models and combines multiple input modalities, including RGB images, depth information, segmentation maps, and text prompts [26]. This indicates a shift from highly specialized systems toward more general counting methods that can operate across a wider range of scenarios and unseen categories.

4.5 Hardware and settings

One of the findings from the initial tests was that the specified working distance functioned relatively well. At this distance, the camera was able to capture images of acceptable quality, which made it possible to identify the material.

The tests revealed some limitations related to image sharpness. Although several adjustments were made to the camera settings, it was difficult to consistently capture sharp images. Different distances between the objects and camera were tested to improve focus and image clarity, but the results did not show any improvement. Even when alternative distances were used, the images were not fully sharp.

Another observation from these tests was the effect of lighting. When the camera was tested close to a window, it became clear that the camera was sensitive to variations in light conditions. Changes in natural light affected image stability and made it more difficult to achieve consistent results. Furthermore, when using the auto exposure mode in the camera settings it caused the camera to respond more slowly to what was happening in front of it. In contrast, the camera performed better in more controlled lighting environments. Tests carried out in the lab and at the station showed more stable results, as both of these locations had uniform industrial lighting. Under these conditions, it was easier to adjust the camera settings and identify a configuration that gave acceptable and consistent images.

The initial tests showed that the camera setup could work for the intended process if the surrounding conditions are controlled. The specified working distance was suitable, but there were still problems with image sharpness. The results also showed that lighting conditions have a strong effect on camera performance. Stable and uniform industrial lighting gave better results than daylight or changing light conditions, which suggests that the camera is likely to work more reliably in the station environment than in less controlled environments.

One example from each class from the dataset containing pen, stripes, and cylinders is shown in Figure 4.8.

(a) Pen

(b) Stripe

(c) Cylinder

Figure 4.8: Example picture of each class: pen, stripe, and cylinder.

The results achieved after labeling and training showed very strong overall performance for the three object classes. Across the evaluation results, the classes were detected with very high accuracy, while the main weakness was related to the background.

The confusion matrix in Figure 4.9 showed that the model classified pen, stripe, and cylinder very well, with almost all predictions being correct. It is important to note that the background class should not be interpreted in the same way as the real object classes. No true background samples were provided in the dataset, and the background class was added automatically by the training and evaluation process. The result presented in Figure 4.9 indicates that the model sometimes confused objects, particularly pens, with the background.

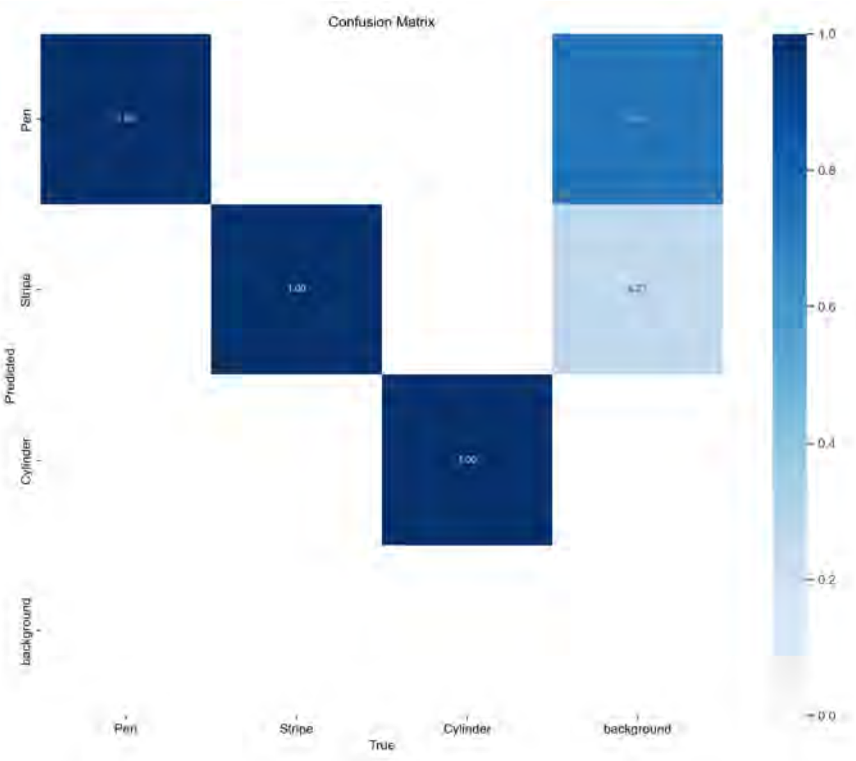


Figure 4.9: Initial test: confusion matrix.

The precision and recall curves (see Figures 4.10 and 4.11) also shows the strong performance of the model. Precision remained high, especially at higher confidence thresholds, which means that the model made very few FP detections. Recall also remained high, meaning that the model was able to detect most target objects. Recall decreased when the confidence threshold became very high, since some correct detections were then excluded. The F1 score, Figure 4.12, showed that the model achieved a good balance between precision and recall. In addition, the precision-recall curve, Figure 4.13, demonstrated strong performance across all three classes. Where the average precision values are close to 1.0. The overall mAP@0.5 was also high, showing that the detector performed reliably.

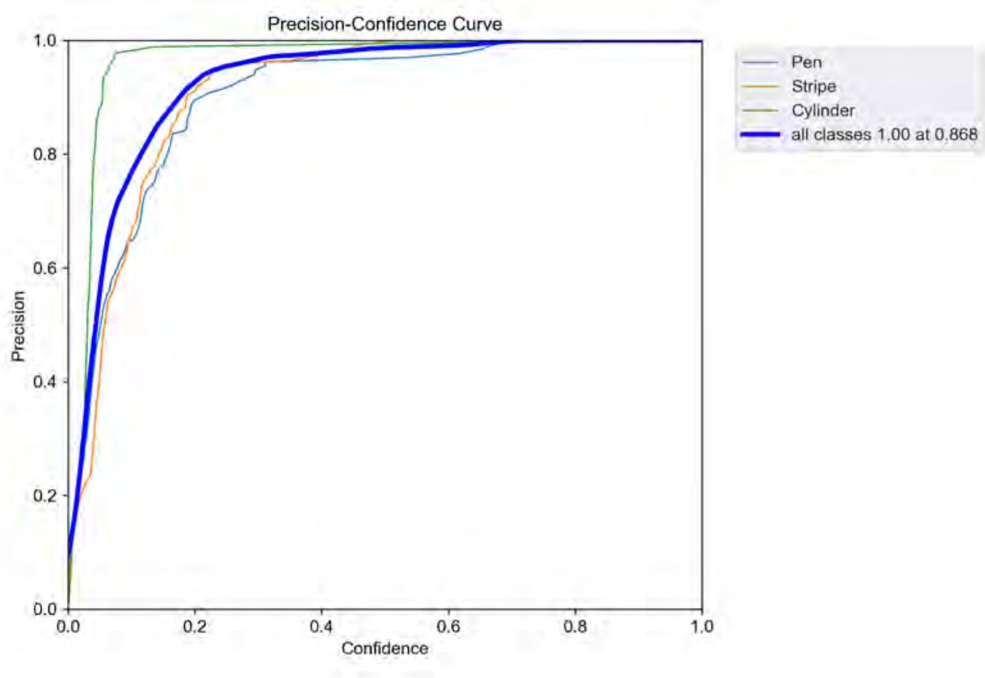


Figure 4.10: Precision curve.

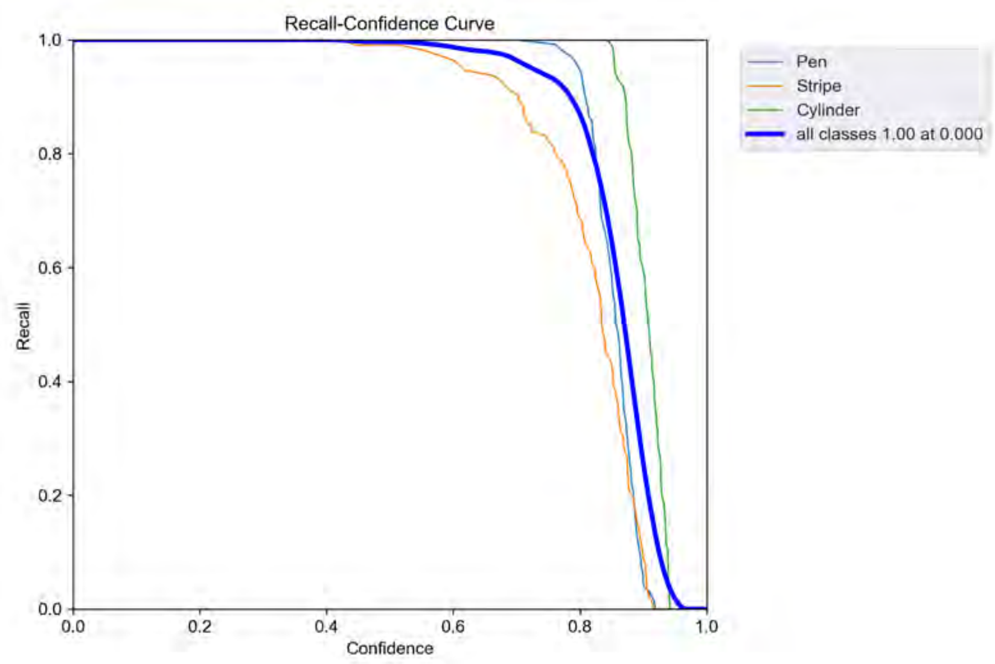


Figure 4.11: Recall curve.

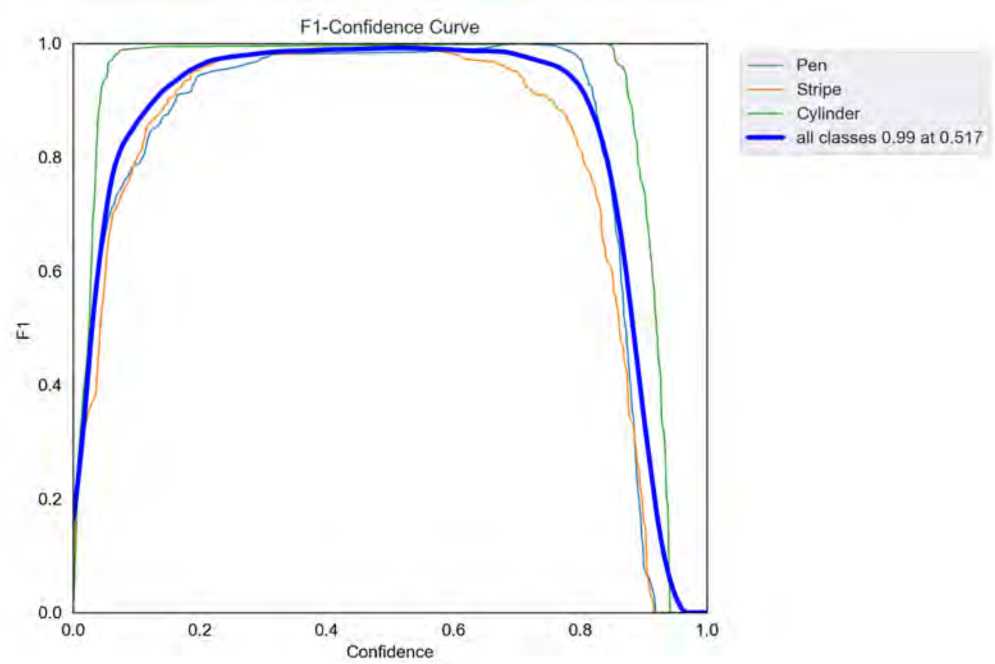


Figure 4.12: F1 curve.

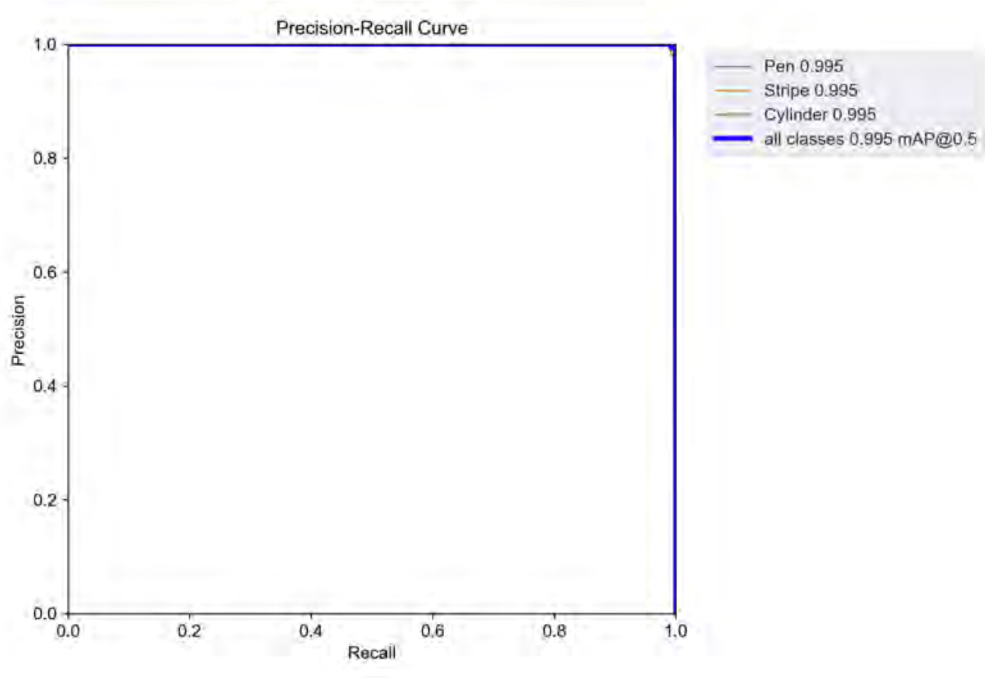


Figure 4.13: Precision-recall curve.

New images collected at the station were used to test the model. These results further showed that the background was not optimal for the trained model. Patterns in the blue box lids, such as circular shapes, were incorrectly detected as cylinders. This can be seen in Figure 4.14.

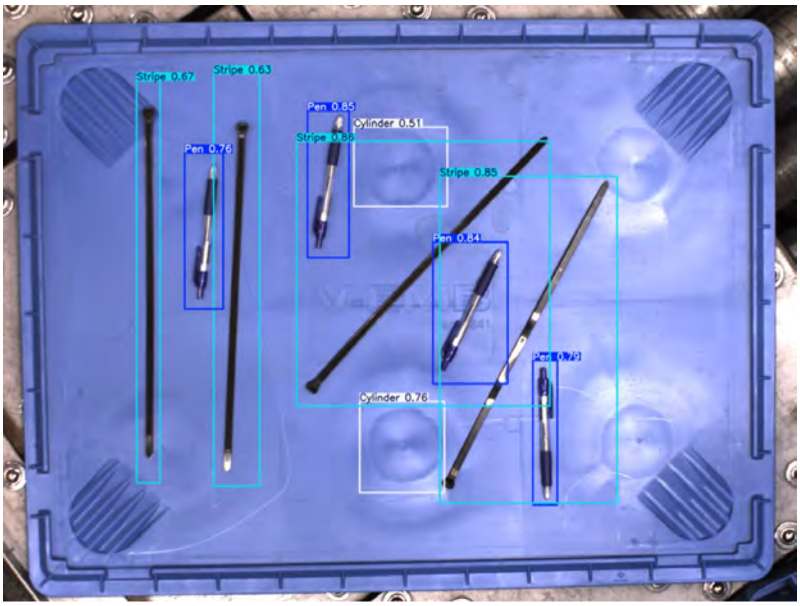


Figure 4.14: Initial test with unseen images from the station.

Based on these results, it was concluded that a different background was needed. A more uniform background was considered important in order to reduce the risk of false detections caused by patterns or variations in the original setup. It was also important that the background did not reflect light, as reflections could affect image quality and make it harder for the model to make correct detections. Therefore, a gray fabric background was selected for the main tests at the station.

4.6 Station concept development

Six station concept ideas were developed based on the defined requirements and the information collected up to this point:

- Manual counting + camera
- Solely computer vision (VVS)
- Solely computer vision (other methods)
- Scale + camera
- Scale + computer vision
- Fully autonomous

All six alternatives were discussed at a conceptual level. However, the greatest emphasis was placed on the solely computer vision (VVS) concept, as this was the main focus of the project.

4.6.1 Manual counting + camera

One of the concept alternatives considered was manual counting combined with a camera system. In this concept, the operator manually counts the material and places it on the packing surface. The operator then triggers the camera, after which an image is automatically captured, renamed, and stored by the system. This means that the solution supports the documentation part of the process, while the counting itself remains manual.

The main advantage of this concept is its simplicity. Compared to a vision-based verification system, it is less complex to implement and therefore less costly. Another benefit is that the image handling process can be automated, which removes manual administrative work related to naming and storing images. In this way, the concept can reduce non value adding activities and improve process efficiency.

The concept also has limitations. Since the counting is still performed manually, the system does not provide direct feedback to the operator during the quantity control task. The camera is only used to store proof of the packed material and does not verify either the quantity or the article type. As a result, the concept does not actively prevent errors, but only documents the process after the operator has completed the task.

This concept can be seen as a simple and cost effective alternative that improves documentation handling, but it does not address the challenge of supporting or automating the quantity verification process.

4.6.2 Solely computer vision (VVS)

The next alternative was solely computer vision based on VVS. In this concept, the system identifies the article and counts the number of objects using a trained object detection model. The concept would automatically store and sort the images similar to the previous concept.

The main advantage of this concept is that it can verify not only the quantity of the material, but also whether the correct article has been packed. In this way, the solution has the potential to provide more direct process support and reduce the risk of packaging errors compared to concepts that only document the packed material.

This concept also comes with challenges. Since this concept relies on trained models, it means that time and effort is required to train the model to recognize all relevant parts. This can be especially hard in this process with many different article types and high variation between materials.

A process map was created to illustrate a possible future state, as shown in Figure 4.15. Compared with the current state process maps in Figures 4.1 to 4.4, the future state would reduce the additional administrative work related to image handling and control.

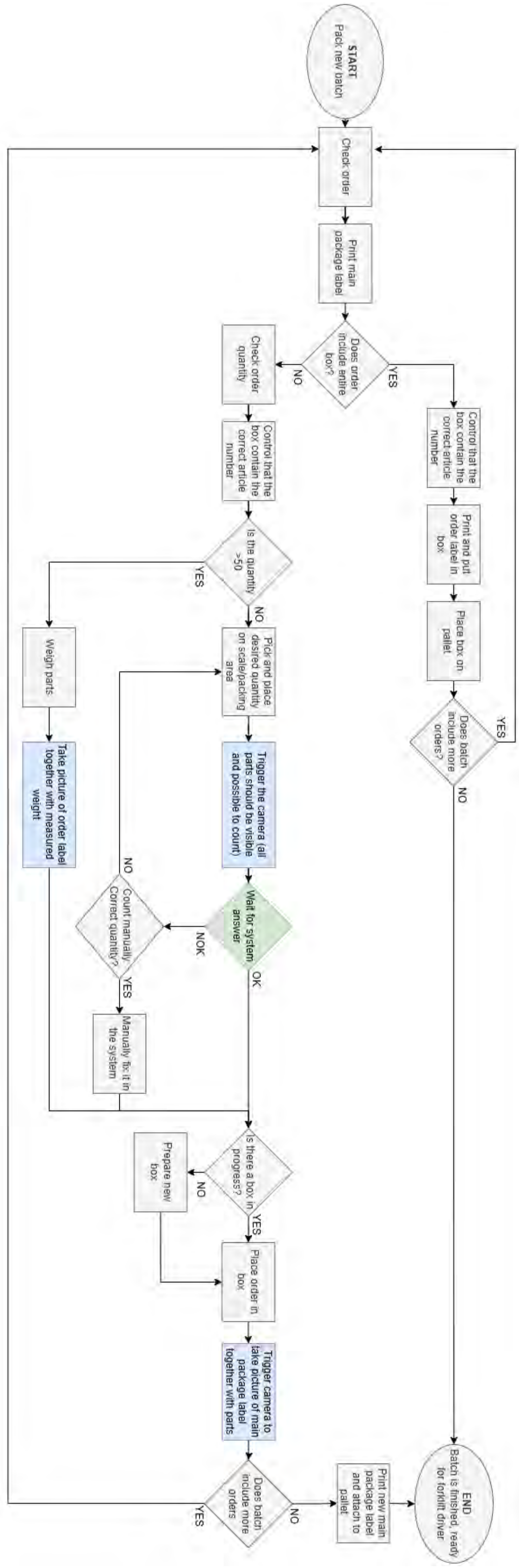


Figure 4.15: Process map: Future state concept.

4.6.3 Solely computer vision (other methods)

Another concept that was considered was solely computer vision using methods other than VVS object detection. This concept is similar to the VVS-based alternative in that it relies on images and computer vision to count the material, but it uses a different detection approach.

One possible approach is segmentation-based counting, where the objects are separated from the background based on differences in features such as brightness, color, or texture. This can work well in controlled environments, especially when the background is uniform and the objects are clearly visible. However, the method can be sensitive to changes in lighting, shadows, and object appearance, which may reduce reliability.

Another possible approach is density map based counting, where the system estimates the number of objects without detecting each instance individually. This method is often useful when many objects overlap or are densely packed. However, it provides less detailed information about the exact location and identity of each object.

A third option is few-shot or exemplar-based counting, where the system is given a small number of examples and then searches for similar objects in the image. This approach can be more flexible than class-based object detection, since it may require less training data for new object types. At the same time, its performance depends on how representative the examples are and how much the target objects vary in appearance.

4.6.4 Scale + camera

The concept scale + camera would be that the quantity is determined by measuring the total weight of the material and dividing it by the weight of each part. Here, the counting would be based on weight rather than computer vision. The camera is instead used to support the documentation process by automatically taking, renaming, and storing images.

A main advantage of this concept is that it does not depend on vision accuracy for counting. This can make the solution easier to implement than a vision-based alternative.It may also be cheaper, require less processing, and respond faster. In addition, the concept can provide direct feedback to the operator, which is good in the packing process.

The concept also has some limitations. Since the counting is based on weight, the result becomes sensitive to variation in part weight. This can be a challenge if the materials differ slightly between individual items. Another limitation is that scale accuracy decreases for smaller and lighter parts, which may reduce counting reliability for certain article types.

4.6.5 Scale + computer vision

Another concept alternative was scale + computer vision. In this concept, the vision system checks the object count, while the scale would contribute as described in previous concept. By combining these two methods, the system receives more information, which could make it easier to identify the object and verify the quantity.

An advantage of this concept is that it may provide higher accuracy than using only one method. The combination of vision and weight data can also reduce the risk of errors caused by overlapping objects or by weight variation in the material.

A limitation could be, as with the scale concept, that the scale accuracy can be lower for smaller parts. In addition, combining two systems increases both the computational time and the complexity of the solution.

4.6.6 Fully autonomous

The final concept was a fully autonomous solution. In this concept, the process would be performed automatically, for example by using a robot system instead of having manual operators.

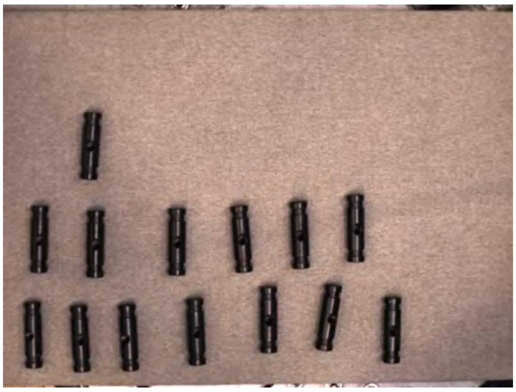
The main advantage of this concept is that it has the potential of higher accuracy while also reducing manual labor. By automating the process, the solution could reduce repetitive tasks and possibly improve process consistency.

However, this concept also has disadvantages. A fully autonomous solution would involve a high implementation cost and would require an external supplier. In addition, the concept is complex.

4.7 Station-based tests

Figure 4.16 shows representative examples from the 11 object classes included in the station-based tests. A total of 100 images were collected for each class, resulting in a dataset containing variations in object orientation and number of objects in the pictures. The object classes differed in size, material, and shape. Manual labeling of the images required approximately 11 working hours to complete.

a) Screw: 0996



b) Pin: 1472

c) Antenna base: 1779

d) Bracket: 3692

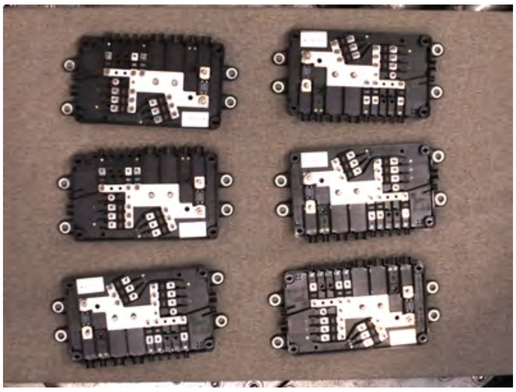
e) Hose: 3897

f) Clamp: 4628

g) Hose clamp: 4733

h) Trim panel: 7433

i) Pipe: 7629



j) Fuse box: 8516

k) Bracket: 9975

Figure 4.16: Articles of different sizes, materials, and shapes used for main tests.

4.7.1 Model development

In this section, the results from training the model using 50, 70, and 90 unique images of each class, with both YOLOv5 Small and YOLOv5 Medium, are presented. Representative result figures are included here, while the complete set of result plots from all training runs can be found in Appendix A.2. Since all experiments showed similar curve trends, this section only uses images from some training runs as examples.

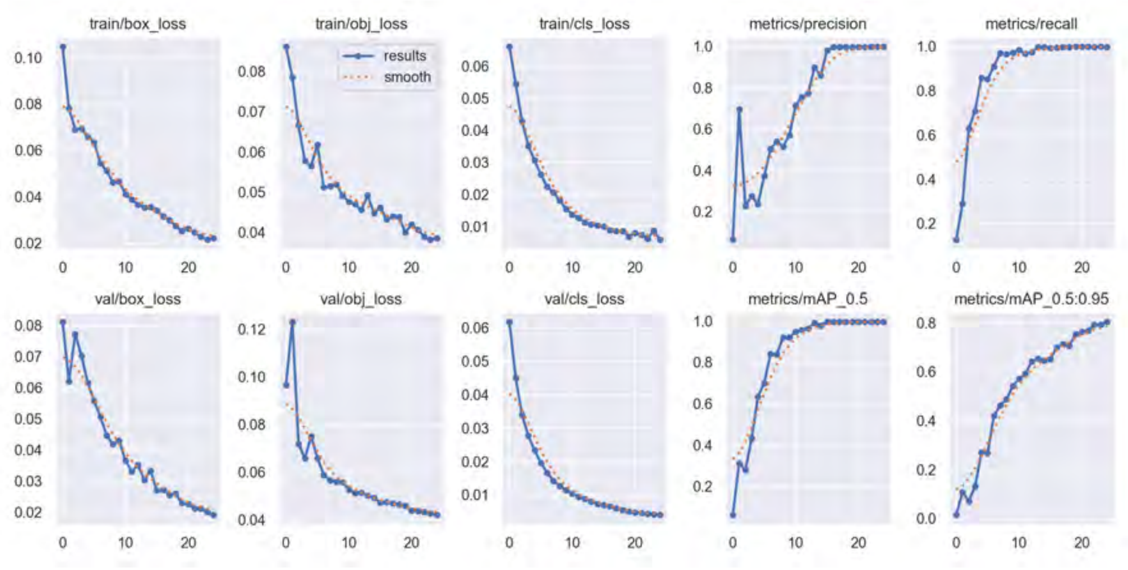
As shown in Table 4.7.1, all tested YOLO models achieved very similar results. The main differences between the models are found in the confidence thresholds. Precision reached 1.00 at thresholds between 0.893 and 0.947, which means that the models needed relatively high confidence scores before all predicted detections were correct. For the F1-score, the maximum value of 1.00 was reached at thresholds between 0.693 and 0.761. Since the F1-score represents the balance between precision and recall, this shows that the models achieved the best trade-off between these two measures at moderately high confidence levels. Recall reached 1.00 at a confidence threshold of 0.000 for all models. This means that when no predictions were filtered out based on confidence, the models detected all objects in the evaluation data. However, this result is expected, since a threshold of 0.000 includes every prediction. For this reason, the recall result is not very informative.

The differences in confidence thresholds were small. This suggests that the models were stable across the different training settings and that increasing the number of training images or using a larger model did not lead to any clear change in detection behavior. There is no clear evidence that the medium model performed better than the small model, or that a larger number of training images improved the final performance.

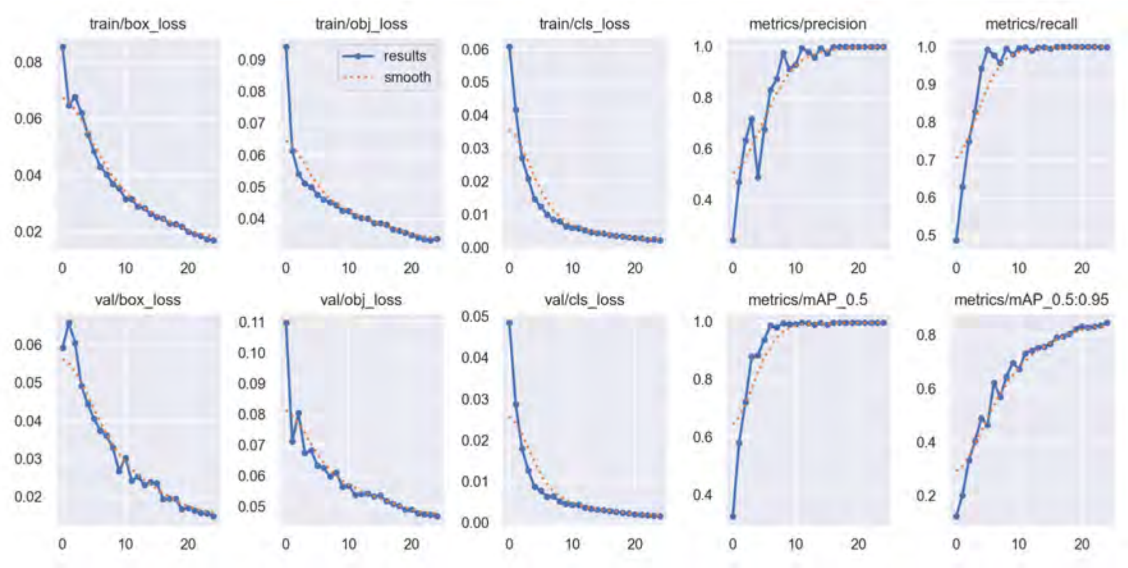
Images	**YOLO model**	**Time (min)**	**Precision at conf.**	**Recall at conf.**	**F1 at conf.**	**mAP@0.5**
50	Small	16.2	1.00 at 0.893	1.00 at 0.000	1.00 at 0.732	0.995
50	Medium	17.1	1.00 at 0.894	1.00 at 0.000	1.00 at 0.693	0.995
70	Small	18.9	1.00 at 0.924	1.00 at 0.000	1.00 at 0.748	0.995
70	Medium	26.3	1.00 at 0.946	1.00 at 0.000	1.00 at 0.759	0.995
90	Small	8.6	1.00 at 0.943	1.00 at 0.000	1.00 at 0.761	0.995
90	Medium	15.3	1.00 at 0.947	1.00 at 0.000	1.00 at 0.758	0.995

The training metrics in Figure 4.17 show that both example models followed a stable learning process. For both of the runs, the training and validation losses decrease clearly during the epochs. This suggests that the models learned the task and that the optimization converged in a stable way. Furthermore, in both cases, precision, recall, and mAP@0.5 increase rapidly during the first part of training and then level off close to 1.00.

A comparison of the two training runs shows very similar curve shapes. The model trained with 90 images per class appears slightly smoother and more stable, but the overall difference is small. This supports the results in Table 4.7.1, where both models achieved very similar performance. Therefore, the training metrics also indicate that increasing the number of training images and using a larger model did not lead to a clear improvement.



(a) 50images/class YOLOv5 Small.



(b) 90images/class YOLOv5 Medium.

Figure 4.17: Main tests: training metrics.

The confusion matrices in Figures 4.18 to 4.20 show a very similar pattern for all the examples presented here. Most classes in all three matrices have values of 1.00 on the main diagonal. This means the model correctly predicted most objects in their true class. These results show that the model can separate the object classes well.

The comparison between the three matrices suggests that the class-level performance is very similar across all tested settings. In all cases, the strongest values are concentrated on the diagonal, while most values that are not on the diagonal or related to the background class are close to zero. This means that classification errors between the object classes are rare. Therefore, the confusion matrices also support that different settings all achieved strong detection and classification performance.

As explained in Section 4.5, the background should not be interpreted the same way as the other classes. The background-related values in Figures 4.18 to 4.20 mainly show unmatched cases during evaluation. In particular, values in the background column can be interpreted as FP detections, meaning that the model predicted an object class but no matching ground-truth object was available. Values in the background row indicate FN, where a real object was present but the model failed to make a detection.

Overall, the comparison suggests that changing the model from small to medium and increasing the number of training images did not lead to large changes in class prediction performance, although Figure 4.20 appears slightly more stable.

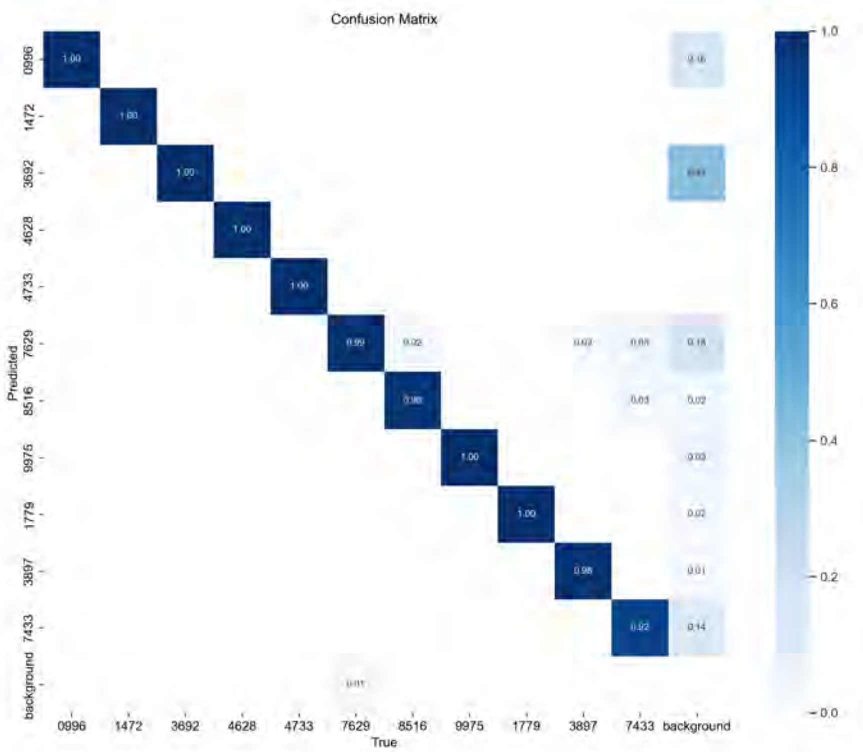


Figure 4.18: 50images/class YOLOv5 Small.

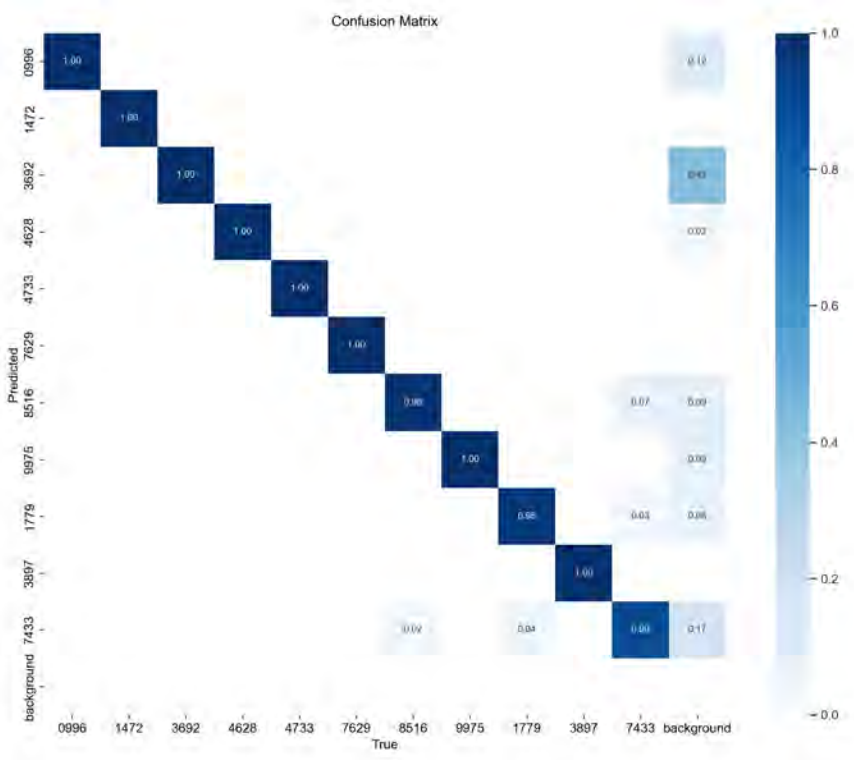


Figure 4.19: 50images/class YOLOv5 Medium.

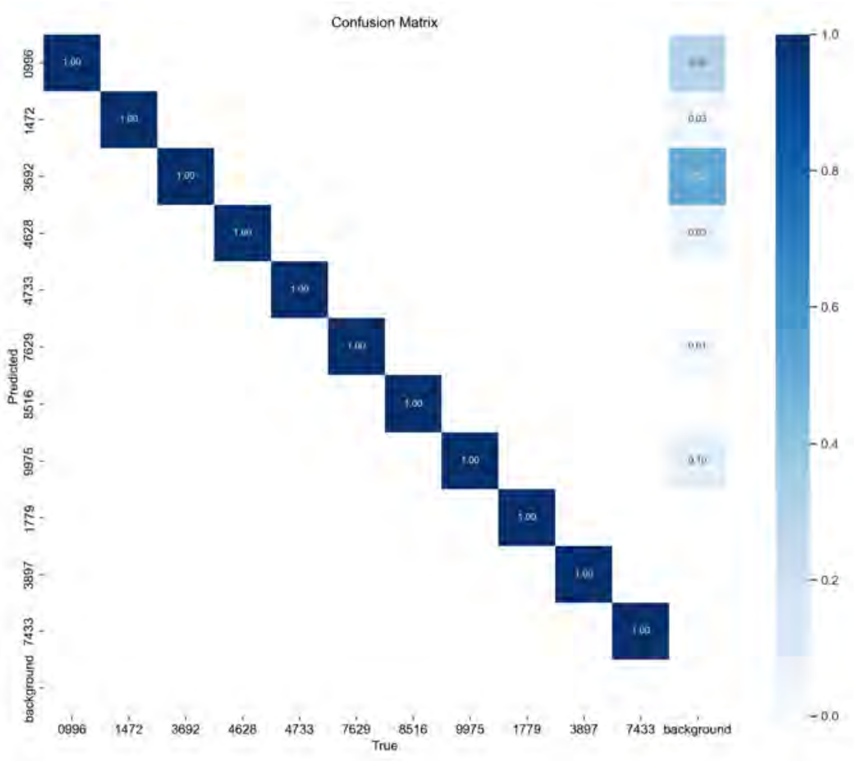


Figure 4.20: 90images/class YOLOv5 Medium.

5

Discussion

This chapter discusses the main findings from the study as well as challenges, possibilities, and practical aspects of implementing a new counting solution at the CKD stations. Additionally, the research methodology followed throughout the study, how this work contributes to already existing and future research, and future development ideas are discussed as well.

5.1 Challenges for industrial implementation

This section discusses the identified challenges, both at the stations but also from a broader point of view, and how those may affect a potential vision-based solution.

5.1.1 Environmental conditions

As the initial tests demonstrated that environmental conditions affect the system performance and accuracy, this was one of the main challenges addressed throughout this project. For instance, lighting and background variations resulted in reduced robustness and FPs, such as the wrongfully detected cylinders on the original blue box lid in Figure 4.14. This suggests that successful implementation of a computer vision solution using object detection, more specifically VVS, relies on the complete station design, including controlled lighting, camera placement, and standardized backgrounds. Even using an accurate YOLOv5 model might result in unreliable results if the lighting, shadows, reflections, or backgrounds vary too much.

Although a more uniformed background proved to reduce the amount of FP predictions, maintaining a stable system with high accuracy is still a challenge. For instance, human error will remain an issue as operators might place parts in a way that makes it hard for the system to count them correctly, e.g. overlapping or too close to each other. This suggests that a vision-based solution would not only require reliable conditions and hardware, but also need to be designed in a way that minimizes the risk of human error.

5.1.2 Article variation and scalability

At the CKD stations, the high number of different parts make it difficult to implement a vision solution, especially with VVS as it is based on YOLOv5. This variation makes collecting data a difficult and time-consuming task as object detection methods, such as YOLOv5, need lots of accurately labeled data to work correctly and achieve reliable results [28]. So, while the initial and main tests demonstrated strong results, they do not reflect the resources required to develop and then maintain the system. Significant time was spent on mounting the test rig, collect images, manually change the part placements, label data, and then develop the model, and still, the tests only cover 11 different parts. This highlights one key issue of implementing VVS at the CKD stations, that is how to do that process for all $\approx$ 1300 parts. Long-term maintenance, updating the dataset, and retraining for new or updated parts would likely become major practical challenges during real implementation.

5.1.3 Long-term maintainability

Achieving good performance does not only rely on the station design and large amount of data, it would also require a long-term maintenance plan for both the system and the actual components. Additionally, issues also arise as the active articles are constantly changing at the CKD stations. That leads to questions such as how to ensure that there are always enough images and labeled data for all parts to avoid reduced performance and who would do that work.

Cleanliness of the station and more specifically keeping the camera and lens clean and free from any dust or other particles that could disturb the vision task is vital [17]. One possible solution to this is to keep the components inside a protective enclosure with only an opening big enough to capture pictures. The type of opening is called an 'optical window' [17]. As this window would be the only surface area that would be affected by the production environment, including dust, temperature changes, etc., the material must be chosen with those aspects in mind.

5.2 Interpretation of the results

This section discusses how the main findings were used to address the RQs.

5.2.1 State of the art in industrial part counting

To address RQ1, the systematic literature review was conducted and time was devoted to searching for similar use cases. The results showed that computer vision is widely used and have been used for many years in industrial environments. Recent developments have driven a shift from more traditional methods toward a greater use of DL-based methods, for example YOLO. Such methods are attractive in industrial environments because they can perform real-time detection while maintaining relatively high accuracy [39]. The literature review also showed that vision-based object counting has reached a relatively high technical maturity in structured industrial environments. Many of the identified studies in these settings reported good performance.

However, despite the research found within computer vision, the authors of this thesis could not find many, if any, studies addressing industrial quantity verification under conditions similar to those at the CKD stations. Most identified studies in industrial settings focused on more controlled settings with lower variation in articles, size, material, and/or placement. In contrast, the CKD environment involves a large variation of articles, non-standardized object placement, different packaging materials, and operator interaction which increase the complexity of using vision. The literature review therefore indicates that there is still a gap in this research area.

The results from this thesis contribute with practical insights into the challenges of implementing computer vision systems in real industrial environments. In particular, the study highlights how environmental conditions and a large number of article types affect the feasibility of industrial part counting systems using computer vision.

Another conclusion that can be drawn from the literature review is that the lack of closely related studies may suggest that using computer vision for counting in industrial settings, such as at the CKD stations, is difficult to implement in practice. Although the technology itself has developed rapidly, real production environments still create challenges that are difficult to fully solve using only a vision-based method.

5.2.2 Feasibility of VVS at the CKD stations

Assessing the test results to answer RQ2, the results indicate that a vision-based solution has potential for industrial part counting and quantity verification at the CKD stations. More specifically, the results showed that VVS could successfully detect and classify several industrial parts under controlled conditions with high accuracy.

The initial tests demonstrated that the system performed well when the environment was more structured, with stable lighting conditions, and a uniform background. Under these conditions, the object detection model achieved reliable detection and classification results. Furthermore, the main results showed that increasing the number of training images led to that the model training and performance metrics became smoother, which is consistent with theory stating that DL-based object detectors require representative datasets to achieve robust performance [35].

At the same time, the results also showed several limitations that affect the feasibility of implementing VVS at the CKD station. One major challenge was the large variation in articles, which complicates both data collection and model training. Transparent plastic bags and reflective surfaces were also problematic since they can affect image quality and object visibility. This supports findings from literature stating that lighting conditions influence computer vision reliability [13].

The results further indicate that environmental stability is important for achieving reliable performance. Small changes in shadows, reflections, and lighting could affect the image acquisition process and thereby influence the detection results. This suggests that a possible future implementation would require a controlled station setup with standardized lighting and placement conditions in order to achieve stable long-term performance.

Although there are currently around 1300 active articles at the CKD stations, many of them are visually similar. This may simplify the implementation process to some extent. Instead of training the model to recognize every individual article separately, it could potentially be trained to recognize more general article categories, such as screws or brackets. Such an approach could reduce the amount of required training data and simplify both labeling and model maintenance. However, this thesis did not evaluate whether this would provide sufficient accuracy for quantity verification in practice.

Another important observation is that the study mainly demonstrates technical feasibility and not long-term robustness. The tests were performed during a limited period of time and under partly controlled conditions. As a result, the study cannot determine how the system performance would change over time in full-scale production. Challenges related to long-term maintainability, introduction of new article types, dataset updates, and system retraining were identified but not fully evaluated.

5.3 Methodological reflection

This section comments on the chosen research methodology as well as factors that may have affected the experimental reliability.

5.3.1 Reflection on following DSRM

By following the structured DSRM process proposed by Peffers et al. [2], there was a clearer connection between all the different phases. This helped ensure that the project was conducted systematically and it also strengthened the academic structure and validity.

One important advantage of following the DSRM in this study was that it allowed for continuous iteration between the activities. In this way, the methodology supported an iterative development process rather than a strictly linear workflow which was really helpful during this project.

However, despite the advantages of following DSRM, some challenges were also identified. In practice, some of the different phases overlapped. For instance, demonstration and evaluation became closely connected during the main tests as the artifact was both demonstrated during the main tests at the station as well as it was continuously evaluated. This made the process more iterative and less sequential than what a first glance at the DSRM model may suggest. However this is also discussed by Peffers et al. [2], where they explain how some previous research suggested a clear distinction between demonstration and evaluation whilst others discussed the phases more closely connected.

As previously described, the first phase of DSRM is the problem identification and motivation [2]. In this project, defining the problem was not as straightforward as initially expected. The existing photo audit process was implemented in order to store proof of shipping the correct quantity, thus reducing the high cost related to customer claims. Insights gained during the interviews and discussions with stakeholders included that there had been several situations in which the stored images indicated that the correct quantity had in fact been sent from Volvo Tuve. This raised questions such as why customer claims still occurred despite the correct quantity being packed at the CKD stations. These uncertainties related to the root cause of the issue made it more complex to clearly describe the problem and formulate the motivation for the project.

Overall, the DSRM approach was considered suitable for this thesis as it supported an iterative work process while providing a systematic framework for developing and evaluating the proposed solution. Moreover, the main challenges encountered during the project were more related to complexity and uncertainty of the industrial context, rather than to the actual methodology.

5.3.2 Experimental uncertainty and limitations

Several factors affected the tests certainty and may have affected the results. To avoid such uncertainties during the testing, the main tests were conducted at the same station each day. Still, complete repeatability was difficult to achieve. For instance, as the test rig was set up for each testing opportunity, small differences in camera placement likely occurred between different tests. Additionally, some camera settings, such as light intake, could be adjusted directly on the actual camera, which introduced the risk of accidentally changing those parameters between tests. Hence, the image quality can potentially vary between the different parts and the performance of the model for each part could potentially be affected as well.

Another important limitation is the limited dataset size and how good it actually represents the real part variation. Due to time limitations and access to different parts at the station, there was no possibility of testing all part types. Furthermore, it was not possible to choose freely what articles to include in the testing. Instead, it was random depending on what parts were available at the stations at the time. This made it hard to ensure a wide variety of parts and the tested articles might not truly represent the extensive part variation.

As the parts included in the station-based tests varied significantly in size, one aspect to consider is the resulting difference in the number of labels per image and per article in total. More specifically, larger objects took up more space in the images, meaning that fewer parts could fit in each photo. This lead to fewer bounding boxes compared to the images of smaller parts, where many objects could appear simultaneously.

Additionally, some parts were excluded from the tests because their physical features and appearance made them specially difficult for a computer vision model to distinguish. This could for example be very long cables, such as the ones in the bottom of Figure 4.5b, that would not be possible to place in a way without them overlapping. It could also be parts that were prepacked in plastic, see Figure 4.6. Excluding those types of articles resulted in more favorable testing conditions and likely higher model performance than what would have been under normal circumstances. So, the performance should not be interpreted as representing the actual expected performance across all article categories.

Another limitation is that there were only two different backgrounds tested, the original blue box lid and the gray fabric. While the gray background reduced the number of FP compared to the blue lid, the results still showed that the main weakness of the model was connected to the background. The limited available backgrounds also means that there is possibly another background better suited for the task. Therefore, additional tests should be conducted to determine what background, is the optimal one. Those tests should cover different materials, textures, colors, and so on.

5.4 Ergonomic factors and ethical considerations

Using computer vision can help reduce the cognitive load on operators. In this case it would support the station operators by confirming that the part quantity is correct. Additionally, by automating the manual admin work, such as image handling and renaming, a lot of cognitive strain would be discarded while the risk of human errors would be greatly reduced.

The design of the system must also consider physical ergonomic aspects. This includes how the proposed system affects operator interaction, workflow, and physical movements at the station. Therefore, when designing the new system it was of great importance to have a solution that could easily be integrated into the current workflow and support the operator without increasing workload or complexity [59].

When introducing new automated systems or solutions in a production environment there are various ethical aspects that must be considered. One aspect in this study is that the proposed solution and the use of computer vision is supposed to support the operators in their work. However, using vision cameras at the CKD stations introduces a risk that it could be perceived as a means to monitor and control the operators. Therefore, it is vital to have clear communication beyond the use of the system and make its purpose clear to both the employees and the management.

One important aspect when designing the new VVS-based station concept was to ensure that no operators would risk being photographed. This is especially important as the images will be saved as proof for a longer period of time. This was done by building a test rig that has a maximum distance between the camera and the packing area that would not allow the surrounding environment to accidentally get in the shot. Moreover, it is also essential that the communication will be continuously clear towards the operators so that they know that there is no risk of being in the image and clear instructions on how to correctly capture the images to not risk any personal data is captured.

Another important issue that arises in projects such as this one is responsibility and accountability. Questions like who is responsible if the system provides the operators with incorrect feedback, for instance, saying that the quantity is correct when it is not. Is the operators still responsible, or is it the system? While in this case Volvo is ultimately responsible, it is important to clearly define the computer vision system as solely a tool and to ensure that the operators knows this.

When using a system to give operator feedback there is a risk of the operators relying too much on the system. This phenomenon is referred to as automation bias [60]. This can lead to the operators trusting the system more than themselves and thereby making their own judgment call regarding the quantity. Moreover, Romeo and Conti [60] explains that the risk of AB increases when it is complicated for the person to verify the system feedback. In this case as the operators are used to the counting process, it was deemed to not be too complicated to verify the feedback. This does not eliminate the risk of automation bias but it might at least reduce it. However, automation bias is an issue that will have to be continuously considered. Therefore, if or when a automized counting method gets introduced to the stations it is vital that it is designed in a way that encourages the operator to stay active and not completely rely on the system.

According to the self-determination theory, people have three psychological needs that are important for motivation and well-being [61]. These include autonomy, relatedness, and competence, and are called ARC in short. Introducing a computer vision based solution may influence both autonomy (the need to feel a sense of choice) and perceived competence (the need to feel effective, skilled, and continually growing) [62]. This is because although the more automated system reduces workload it might actually reduce the operator's sense of control over the process and might, as previously mentioned, instead introduce a feeling of being monitored. As for the competence aspect it is important to ensure that the station operators will not feel as though their possibility to grow in their work decreases. In other words, introducing a system that takes over key work tasks there could be a risk that the operators would feel as though the need for their skill diminishes too much. In this project, this created an important consideration to keep a balance between efficiency and meaningful operator involvement. Ensuring that the system supports rather than replaces operator skills was therefore an important aspect when designing the proposed solution and for future implementation. Additionally, ensuring that the operators recognize that they are still in control over the work process at the CKD stations and that the system is only there for support and to provide instant feedback.

5.5 Future development

This section discusses possible improvement and changes, alternative solution ideas, and other future research areas.

5.5.1 Alternative hardware configurations

The camera used for this project is, as previously mentioned, a color camera as it was initially assumed that it would be necessary to being able to distinguish between different colors in the images. However, a monochrome camera might have been sufficient and even the better choice because of the higher quality images. Therefore, if Volvo still chooses to try and implement a VVS solution at the station or for other similar tasks, it should be tested if a monochrome camera would achieve better results.

One aspect that this study could not cover is the light changes throughout the year. One conclusion from the pilot test was that the camera is affected by changes in lightning. There is a window above the CKD stations that during the time of the study did not let in much sun but during the summer it usually lets in a lot of sun. A potential solution could be to use a light to avoid shadows from operators and to maintain stable lighting conditions despite the window. Also to ensure consistent lighting for a more stable and accurate process. This is something that must be studied more closely if implementing an industrial camera at the stations.

5.5.2 Alternative solution ideas

If an industrial camera would be installed at the stations, with regards to all the environmental constraints, the operators could use it to easily and semi-automatically capture images continuously. This would be possible if there was a system that would do the manual image admin work, then eventually enough data could be captured without dedicating too much time or resources on it. This would simplify the work process for the operators while avoiding the challenges related to a complete computer vision system.

Additionally, this would enable an easy way to collect data for a possible future implementation of VVS at the stations. However, it is not a complete solution as it does not provide any feedback to the operators. Still, this might be a good and rather simple solution to implement to keep the photo audit while reducing the need for an operator to do the manual admin work. Thus, freeing up time for the operators to do value-adding work instead. As having an operator costs 855,184 SEK/year this would result in large savings and less waste.

A possible alternative solution is to combine a scale with a simple camera setup. First, one part is weighed to get a reference weight. When parts are picked, the box or container is placed on the scale and the system calculates how many parts have been picked by comparing the total weight with the reference weight. To support traceability, a fixed camera at each station can, as previous solution, take a picture and store it.

5.5.3 Alternative detection methods

As described in chapter 2.2, different detection methods affects the outcome and the performance of the system. Since VVS is based on YOLOv5, i.e. object detection, [24], alternative detection methods could not be tested in practice during this project. Consequently, a complete comparison between different methods could not be performed. If Volvo chooses to look further into implementing computer vision at the CKD stations, it is recommended to do practical tests of other methods before developing a final model.

Object detection models requires a lot of preprocessing, in the form of labeling with bounding boxes [24]. Therefore, for cases such as this where there is no real need to be able to distinguish between different parts, using object detection can mean added, unnecessary, and time-consuming work. Especially when there are a lot of different parts.

Instead, a class-agnostic method, such as few-shot, could be a better alternative. As it only needs a one or a few examples in order to recognize the object [28]. This would mean a lot of time saved for a project such as this with a high number of different parts as preprocessing the data can be very time-consuming [54]. Additionally, the maintainability of the system could be improved as well since changes of part geometry or new parts added would not need to be labeled in the same way as, for instance, object detection would require. Moreover, at the CKD stations there is not really a requirement that the system should be able to know exactly what the object is, but rather count the parts and provide the operator with feedback. Further suggesting that using a category specific method for this project could mean redundant work. However, without any tests at the stations it is not possible to verify if the performance and accuracy of a class-agnostic method would be good enough to be implemented.

6 Conclusion

The purpose of this thesis was to better understand the potential and limitations of computer vision-based part counting at CKD stations. The study investigated the current state of the art in industrial object counting and evaluated the performance of VVS at a CKD station.

The findings from the literature review showed that computer vision is increasingly used in manufacturing and logistics for inspection, object detection, and automation tasks. Different counting approaches exist, including object detection, segmentation-based counting, density-based counting, and few shot methods. The review also showed that industrial environments create challenges related to scalability, object variation, lighting conditions, and system maintenance.

The experimental results demonstrated that VVS using YOLOv5 can successfully detect and count industrial parts in controlled environments. The system performed better when trained with larger datasets, and the results indicate that computer vision-based counting can support quantity verification processes at CKD stations. At the same time, the study identified several limitations. Detection performance was affected by environmental conditions such as lightning and reflections. Furthermore, the collection and preprocessing of the data were time consuming. These factors make industrial implementation more complex than laboratory testing.

The thesis also showed that implementing a vision-based counting system is not only a technical challenge, but also an organizational and operational one. Factors such as station layout, operator interaction, maintainability, and long-term flexibility are important for creating a reliable and sustainable solution. Even though the proposed concept was limited to one station and one industrial case, the findings provide useful insights into how computer vision can be integrated into industrial logistics processes.

Overall, the study concludes that VVS has high potential and performance when it comes to object detection. A computer vision-based solution could reduce manual workload, improve traceability, and create more standardized documentation of packed material. However, in cases such as this one where there is a large amount of highly varied products and manual work, there are a lot of challenges that hinders a smooth implementation of a vision system.

Future work should focus on evaluating the system with a larger variety of parts and under longer operating periods in the real production environment. Further studies could also investigate alternative counting approaches, improved lighting solutions, automated image capture setups, and class-agnostic methods that may increase flexibility when handling large numbers of different article types.

Declaration of AI usage

AI-based tools, including Volvo GPT and ChatGPT, were used during the writing process of this thesis. The tools were primarily used to support language refinement, including improving sentence structure, enhancing clarity, and making the text more concise.

Bibliography

[1] Volvo Group, "Volvo FM in Malaysia." https://media.volvogroup.com/dam/contentitems/558b6b78d36944dca28eb3b5008f9b89, 2025.

[2] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007.

[3] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021.

[4] "Industrial usb3 cameras for industrial and embedded applications," 2026. [Online]. Available: https://www.stemmer-imaging.com/en/av-alvium-1800u500cchc (accessed on: 2026-02-11).

[5] "Machine vision lenses hf-xa-5m series," 2026. [Online]. Available: https://www.fujifilm.com/se/en/business/optical-devices/mvlens/hfxa5m#HF02 (accessed on: 2026-02-09).

[6] M. Bellgran and K. Säfsten, "Design and evaluation of production systems," in *Production Development: Design and Operation of Production Systems*, ch. 8, Springer London, 1 ed., 2010.

[7] R. E. Amtower, "Low cost machine vision systems extend manufacturing applications," in *Proceedings of WESCON 1993 Conference Record*, (San Francisco, CA, USA), pp. 116–119, WESCON 1993 Conference, Institute of Electrical and Electronics Engineers Inc., 1993.

[8] Q. Cheng, "The application of computer vision technology in the field of industrial automation," in *2021 International Conference on Applications and Techniques in Cyber Intelligence* (J. Abawajy, Z. Xu, M. Atiquzzaman, and X. Zhang, eds.), (Cham), pp. 274–279, Springer International Publishing, 2021.

[9] B. Sevilmis, B. Onle, and O. Ozkan, “Density-based counting in cluttered scenes via synthetic data and deep vision models,” *International Conference on Computer Science and Engineering, UBMK*, no. 2025, p. 997 – 1002, 2025.

[10] W. Vexo, C. Jeenanunta, K. Chou, A. Tunpan, N. Sirimarnkit, T. Vong, and K. Thourn, “Real-time image processing for production tracking in manufacturing plant,” p. 77 – 82, Institute of Electrical and Electronics Engineers Inc., 2023.

[11] N. K. Verma, T. Sharma, R. K. Sevakula, and A. Salour, “Vision based object counting using speeded up robust features for inventory control,” p. 709 – 714, 2017.

[12] L. Boone, “Industry 4.0 (fourth industrial revolution),” *Salem Press Encyclopedia*, 2025.

[13] B. G. Batchelor and P. F. Whelan, *Machine Vision for Industrial Appliactions*, pp. 3–59. London: Springer London, 2012.

[14] S. Guo and B. Yu, “Research status and trends of industrial automation based on computer vision technology,” pp. 363 – 368, 2025.

[15] H. S. S. D. H. S. Gill, Rupali, “10. understanding computer vision for intelligent autonomous systems,” 2025.

[16] B. G. Batchelor, *Selecting Cameras for Machine Vision*, pp. 477–506. London: Springer London, 2012.

[17] B. G. Batchelor, *Light and Optics*, pp. 157–258. London: Springer London, 2012.

[18] B. G. Batchelor, *Cameras*, pp. 358–476. London: Springer London, 2012.

[19] B. G. Batchelor and P. F. Whelan, *Human and Animal Vision*, pp. 89–115. London: Springer London, 2012.

[20] “Industrial camera solutions for machine vision,” 2026. [Online]. Available: `https://www.stemmer-imaging.com/en/products/cameras` (accessed on: 2026-05-18).

[21] S. Harvey and J. Bähr, “A temperature-controlled complementary metal oxide semiconductor camera,” *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 218, no. 9, p. 1217 – 1222, 2004.

[22] B. G. Batchelor, *Inspecting Natural and Other Variable Objects*, pp. 61–87. London: Springer London, 2012.

[23] B. G. Batchelor, *Illumination Sources*, pp. 283–317. London: Springer London, 2012.

[24] V. Pagire, M. Chavali, and A. Kale, “A comprehensive review of object detection with traditional and deep learning methods,” *Signal Processing*, vol. 237, p. 110075, 2025.

[25] E. DAVIES, “4 - thresholding techniques,” in *Machine Vision* (E. DAVIES, ed.), Microelectronics and Signal Processing, pp. 77–99, Academic Press, 1990.

[26] Z. Shi, R. Liu, J. Takahashi, and S. Jiang, “Truecount: Improving open-world object counting with visual-language models and dynamic multi-modal inputs,” in *Proceedings of the 33rd ACM International Conference on Multimedia*, MM ’25, (New York, NY, USA), p. 1764–1773, Association for Computing Machinery, 2025.

[27] Z. You, K. Yang, W. Luo, X. Lu, L. Cui, and X. Le, “Few-shot object counting with similarity-aware feature enhancement,” p. 6304 – 6313, Institute of Electrical and Electronics Engineers Inc., 2023.

[28] M. Shi, H. Lu, C. Feng, C. Liu, and Z. Cao, “Represent, compare, and learn: A similarity-aware framework for class-agnostic counting,” vol. 2022-June, p. 9519 – 9528, 2022.

[29] K. Kumar, P. Kumar, V. Kshirsagar, R. H. Bhalerao, K. Shah, P. K. Vaidhya, and S. K. Panda, “A real-time object counting and collecting device for industrial automation process using machine vision,” *IEEE Sensors Journal*, vol. 23, no. 12, p. 13052 – 13059, 2023.

[30] Z. Cui, H. Mei, W. Dong, Z. Wei, Z. Lv, D. Zhou, and X. Yang, “Steel sheet counting from an image with a two-stream network,” *IEEE Transactions on Instrumentation and Measurement*, vol. 74, 2025.

[31] T. Bagyammal, K. Mahesh Madhav, Y. Dhiraj Kumar, N. Arthi Sri, and V. Siva Srinivas, “Object detection in shelf image of the retail store using single stage detector,” p. 218 – 224, Institute of Electrical and Electronics Engineers Inc., 2024.

[32] A. Sharma, Z. Zhang, and R. Rai, “The interpretive model of manufacturing: a theoretical framework and research agenda for machine learning in manufacturing,” *International Journal of Production Research*, vol. 59, no. 16, pp. 4960–4994, 2021.

[33] A. Mahajan, G. Singh, S. Devgan, G. Singh, and P. Kaur, “Next-generation manufacturing: leveraging ai for industrial innovation and growth,” *Computers Industrial Engineering*, vol. 211, p. 111618, 2026.

[34] C. Clavero, M. A. Patricio, J. García, and J. M. Molina, “Towards smarter warehouse perception: integrating adaptive tile segmentation in warehousing object detection pipelines1,” *Expert Systems with Applications*, vol. 299, p. 130003, 2026.

[35] K.-Y. Huang, L. Zou, C.-L. Chen, Y.-T. Ta, I.-B. Liao, Y.-H. Li, and W.-H. Cheng, "C3l: Class-centric contrastive learning for long-tailed learning," in *2025 6th International Conference on Information Technology and Education Technology (ITET)*, pp. 41–45, 2025.

[36] S. V. Carata, M. Ghenescu, and R. Mihaescu, "Real-time detection of unrecognized objects in logistics warehouses using semantic segmentation," *Mathematics*, vol. 11, no. 11, 2023.

[37] C. M. Bishop, *Pattern Recognition and Machine Learning.* Information Science and Statistics, New York, NY: Springer, 2006.

[38] K. Bogomasov and S. Conrad, "Efficient iterative semi-supervised object detection training for counting of fruits and vegetables," in *Proceedings of the 2023 7th International Conference on Advances in Artificial Intelligence*, (New York, NY, USA), p. 133–137, Association for Computing Machinery, 2024.

[39] Ultralytics, "Ultralytics yolo documentation." `https://docs.ultralytics.com/`, 2024. Accessed: 2026-04-23.

[40] J. Yin and J. Jin, "A face mask detection algorithm based on yolov5," in *2022 International Conference on Machine Learning, Control, and Robotics (MLCR)*, pp. 55–61, 2022.

[41] Microsoft, "What is azure?."

[42] Microsoft, "What is azure machine learning?."

[43] Y. Zheng, "Do information systems today play a strategic role in business?," *Journal of Soft Computing and Decision Analytics*, vol. 2, no. 1, pp. 65–70, 2024.

[44] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *Management Information Systems Quarterly*, vol. 28, pp. 75–106, 03 2004.

[45] T. Netland, D. Powell, and P. Hines, "Demystifying lean leadership," *International Journal of Lean Six Sigma*, vol. ahead-of-print, 08 2019.

[46] A. Elhami and B. Khoshnevisan, "Conducting an interview in qualitative research: The modus operandi," vol. 46, p. 2022, 01 2022.

[47] S. Zimmer, "Stakeholder theory and analysis," *Salem Press Encyclopedia*, 2021.

[48] O. A. Adeoye-Olatunde and N. L. Olenik, "Research and scholarly methods: Semi-structured interviews," *JACCP: JOURNAL OF THE AMERICAN COLLEGE OF CLINICAL PHARMACY*, vol. 4, no. 10, pp. 1358–1367, 2021.

[49] B. Kitchenham, "Procedures for performing systematic reviews," Joint Technical Report TR/SE-0401 and 0400011T.1, Keele University and National ICT Australia Ltd., Keele, Staffs ST5 5BG, UK and Eveleigh, NSW 1430, Australia,

July 2004. Keele University Technical Report TR/SE-0401; NICTA Technical Report 0400011T.1.

[50] C.-W. Chou and Y.-T. Hsu, "Robust real-time object detection and counting system for casting foundries," *Applied Soft Computing*, vol. 176, p. 113155, 2025.

[51] H. Zang, Y. Peng, M. Zhou, G. Li, G. Zheng, and H. Shen, "Automatic detection and counting of wheat spike based on dmseg-count," *Scientific Reports*, vol. 14, no. 1, 2024.

[52] S. B. Patil, J. C. Shimpi, A. G. Tanawade, P. G. Chavan, and V. S. Tandulkar, "Autonomous object detection and counting using edge detection and image processing algorithms," p. 1280 – 1286, Institute of Electrical and Electronics Engineers Inc., 2023.

[53] Y. Huang, V. Ranjan, and M. Hoai, "Interactive class-agnostic object counting," p. 22255 – 22265, Institute of Electrical and Electronics Engineers Inc., 2023.

[54] K. Bogomasov and S. Conrad, "Efficient fruit and vegetable classification and counting for retail applications using deep learning," p. 65 – 71, 2021.

[55] E. Omeragic and E. Sokic, "Counting rectangular objects on conveyors using machine vision," 2020.

[56] Q. Zhao, I. Taniguchi, M. Nakamura, and T. Onoye, "An adaptive parts counting method based on orthogonal intensity distribution analysis for industrial vision systems," p. 118 – 119, 2019.

[57] M. Marsden, K. McGuinness, S. Little, C. E. Keogh, and N. E. O'Connor, "People, penguins and petri dishes: Adapting object counting models to new visual domains and object types without forgetting," p. 8070 – 8079, 2018.

[58] N. K. Verma, A. Goyal, A. Chaman, R. K. Sevakula, and A. Salour, "Template matching for inventory management using fuzzy color histogram and spatial filters," p. 317 – 322, 2015.

[59] C. Berlin and C. Adams, *Production Ergonomics*. London: Ubiquity Press, Jun 2017.

[60] G. Romeo and D. Conti, "Exploring automation bias in human–ai collaboration: a review and implications for explainable ai," *AI & Society*, vol. 41, pp. 259–278, 2026.

[61] E. Deci, A. Olafsen, and R. Ryan, "Self-determination theory in work organizations: The state of a science," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 4, pp. 19–43, 04 2017.

[62] S. Fowler, *Why Motivating People Doesn't Work . . . and What Does: The New Science of Leading, Energizing, and Engaging.* Berrett-Koehler Publishers, 2014. Chapter 2.

A

Appendix 1

A.1 Interview structure

The following questions were used as a guide during the semi-structured interviews.

Background and role

1. What is your current role or title?

2. How are you connected to the CKD stations or the implementation of vision systems at the stations?

3. Do you have any previous experience working with vision systems or computer vision?

Business impact and motivation

4. Why is the current picture-taking process necessary? What value does it provide for Volvo?

5. From your perspective, what is the main problem that creates the need for a new solution?

6. Before the current process was introduced, how common were customer complaints related to incorrect delivery quantities?

7. How severe would you say that the issue of incorrect quantities and customer claims is?

8. For how many customers are pictures currently taken?

9. Is it always the same customers that the operators take pictures for? If so, why and how did you choose those customers?

10. Does the type of receiving factory or business model affect whether pictures

are taken?

Statistics and cost-related questions

11. What is the approximate cost of a customer complaint related to incorrect quantity?

12. What are the costs associated with the current picture-taking process?

13. Do you believe the current process results in overall cost savings for the company?

14. What are the costs associated with sending replacement parts after customer complaints?

15. For how long are the images supposed to be stored? How long are the images typically stored in reality?

16. How long do customers have to make a claim?

17. Can you think of any additional data or statistics that could be valuable for this study?

CKD station specifics

18. How frequently is the picture-taking process performed?

19. Approximately how many products are processed at the stations per day or week?

20. How are new employees trained to perform the picture taking process?

21. Are there any standard operating procedures (SOPs), instructions, or other guidelines available for the operators?

22. Approximately how many different article types are included in the picture taking process?

Future state and other thoughts and ideas

23. Why do you believe a vision-based solution could be suitable for this application? Why have not other solutions, such as using a scale, been considered?

24. Do you have any additional ideas, thoughts, or concerns regarding the process or possible future solutions?

A.2 Model development results

A.2.1 Model trained with 50 unique images of each class

YOLOv5 Small

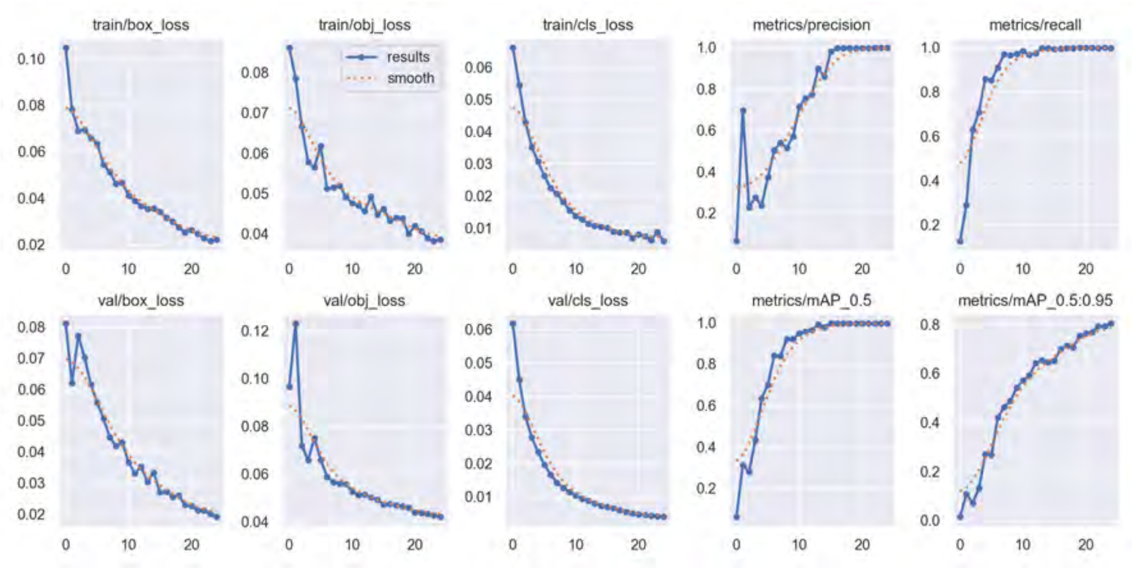


Figure A.1: 50images/class YOLOv5 Small: training metrics

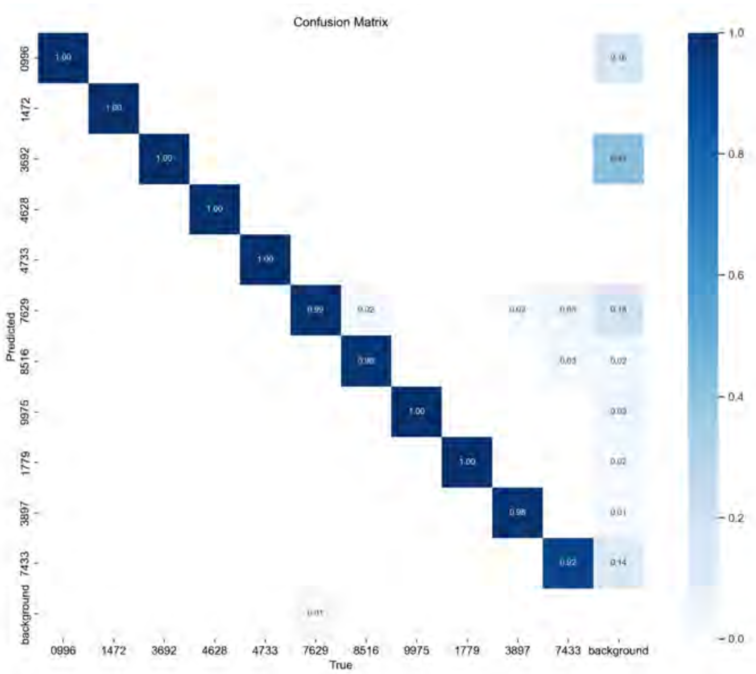


Figure A.2: 50images/class YOLOv5 Small: confusion matrix

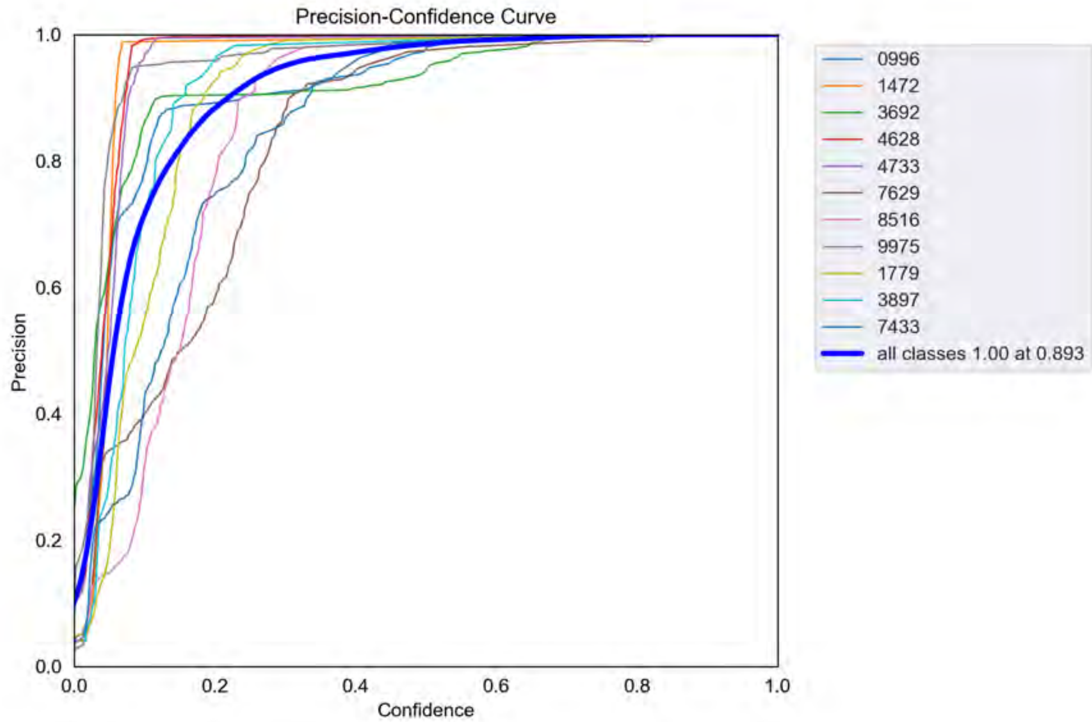


Figure A.3: 50images/class YOLOv5 Small: precision curve

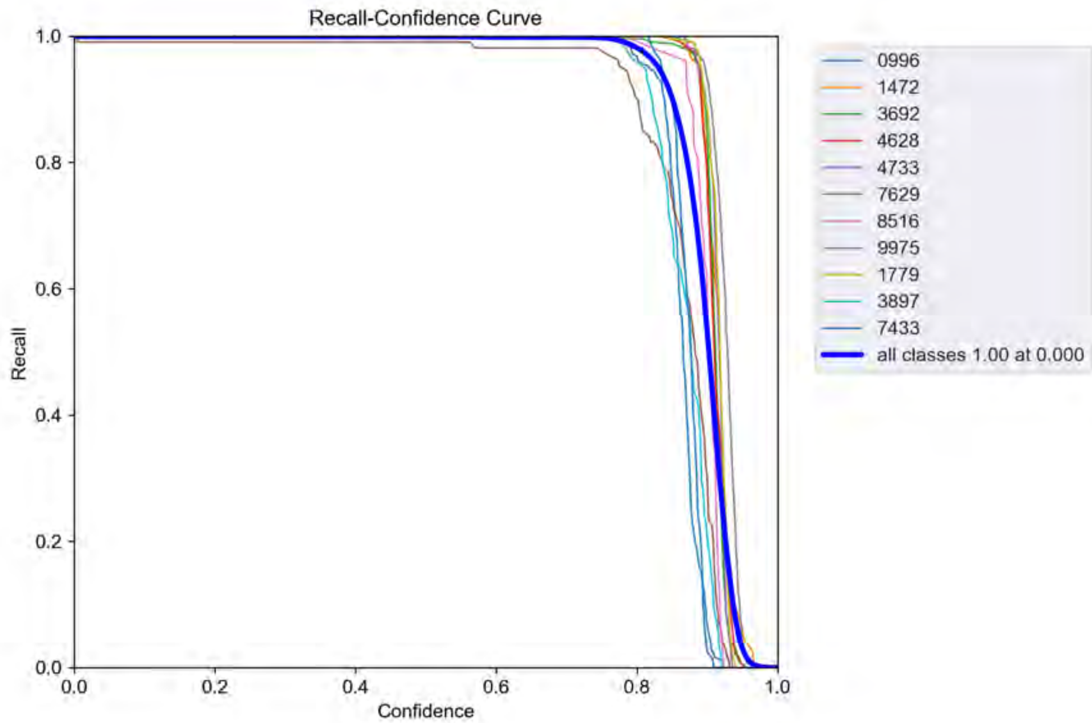


Figure A.4: 50images/class YOLOv5 Small: recall curve

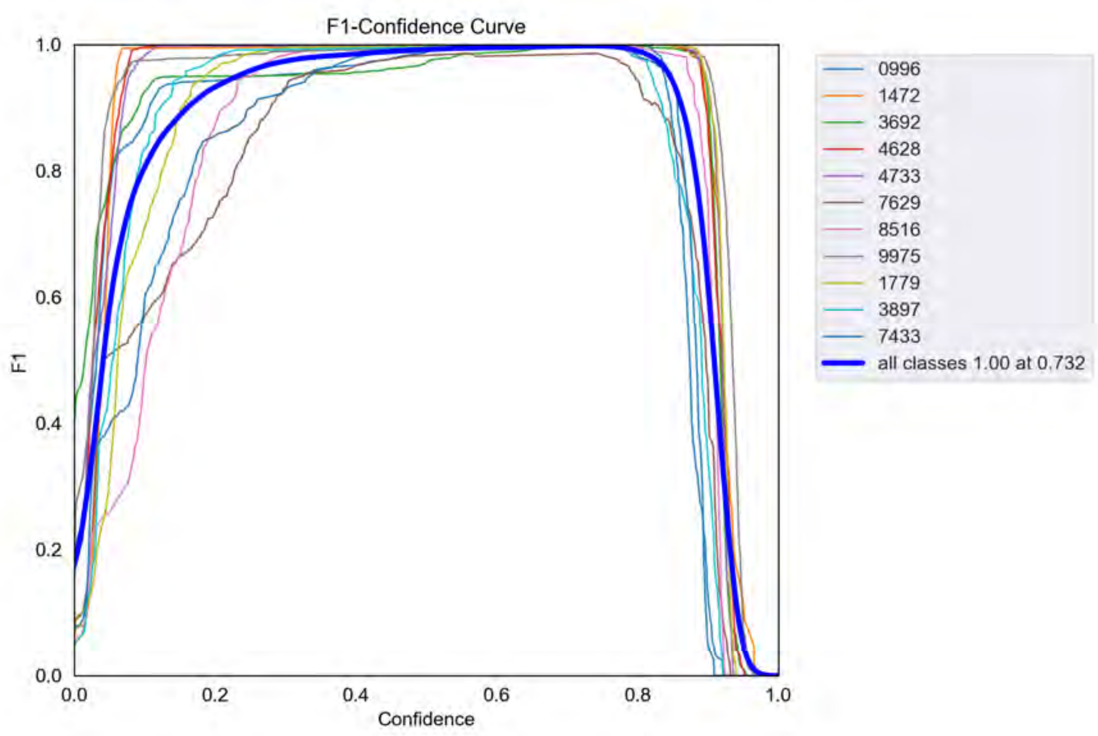


Figure A.5: 50images/class YOLOv5 Small: F1 curve

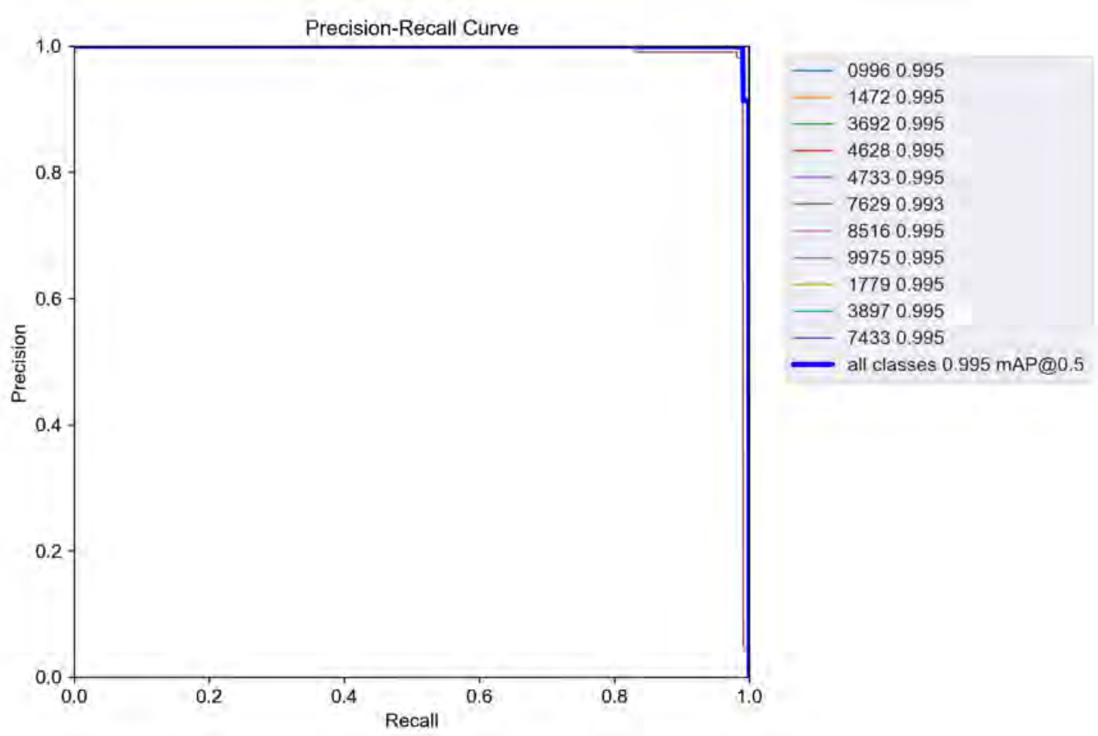


Figure A.6: 50images/class YOLOv5 Small: precision-recall curve

YOLOv5 Medium

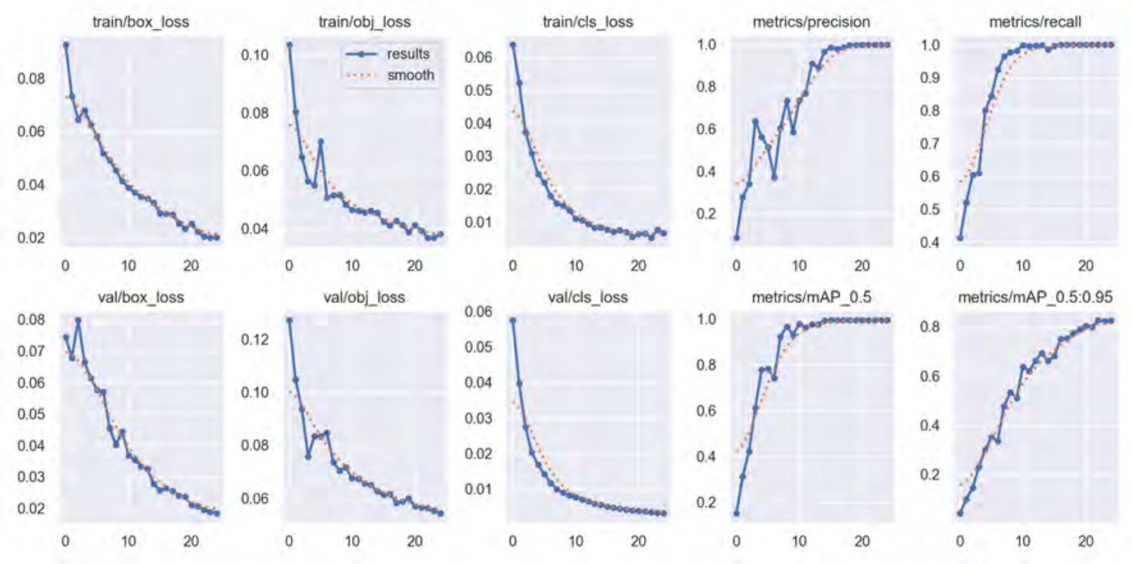


Figure A.7: 50images/class YOLOv5 Medium: training metrics

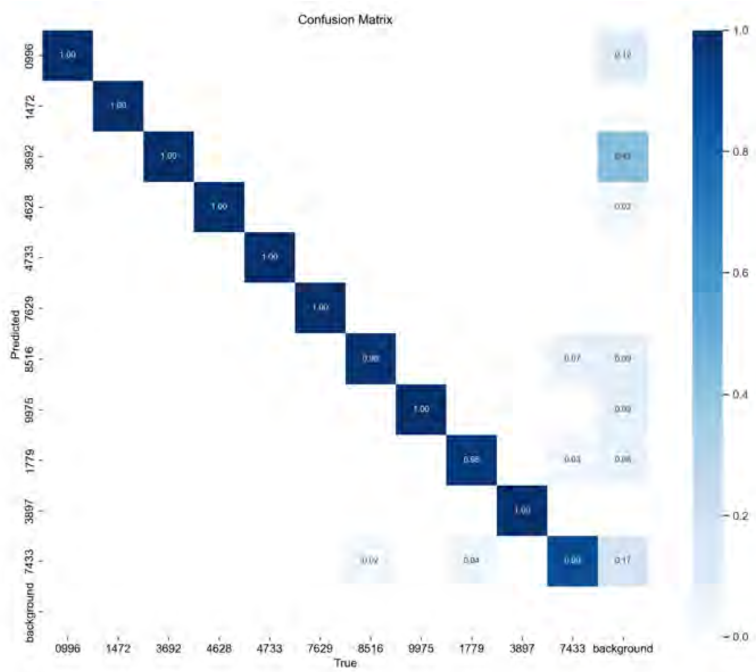


Figure A.8: 50images/class YOLOv5 Medium: confusion matrix

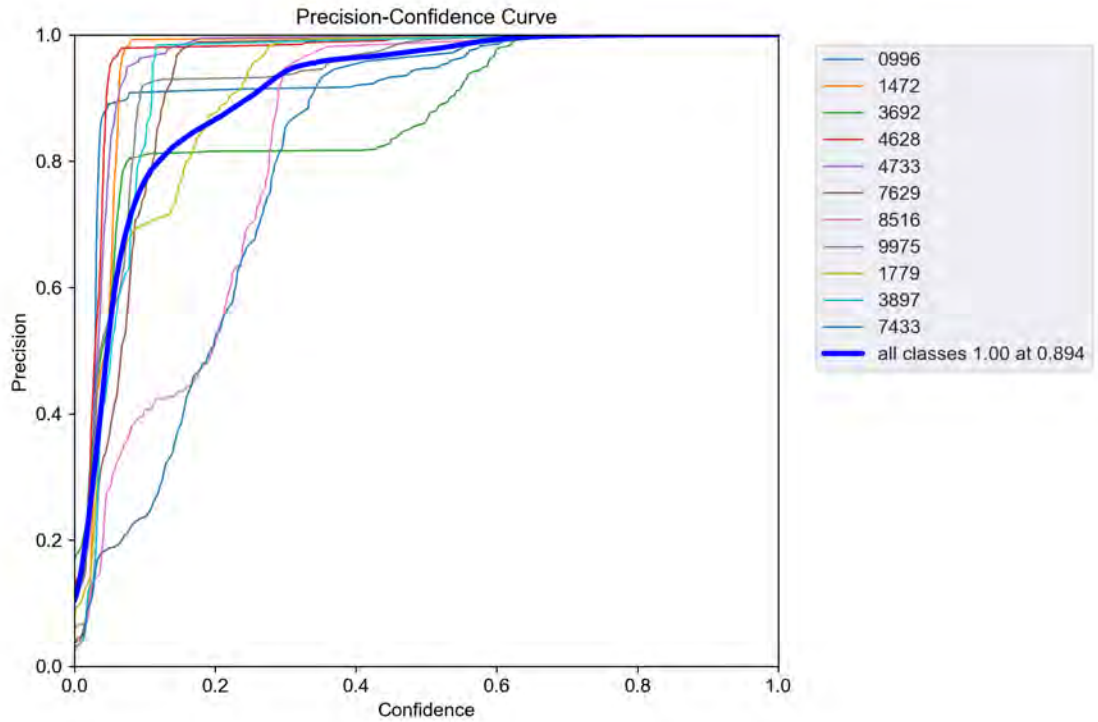


Figure A.9: 50images/class YOLOv5 Medium: precision curve

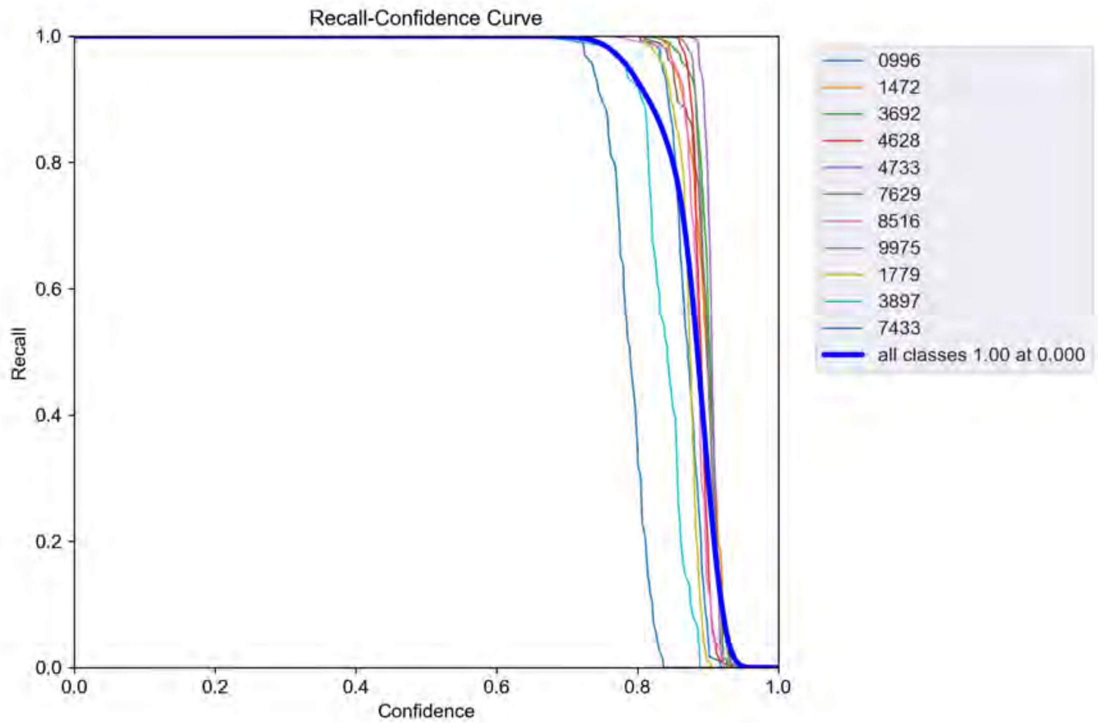


Figure A.10: 50images/class YOLOv5 Medium: recall curve

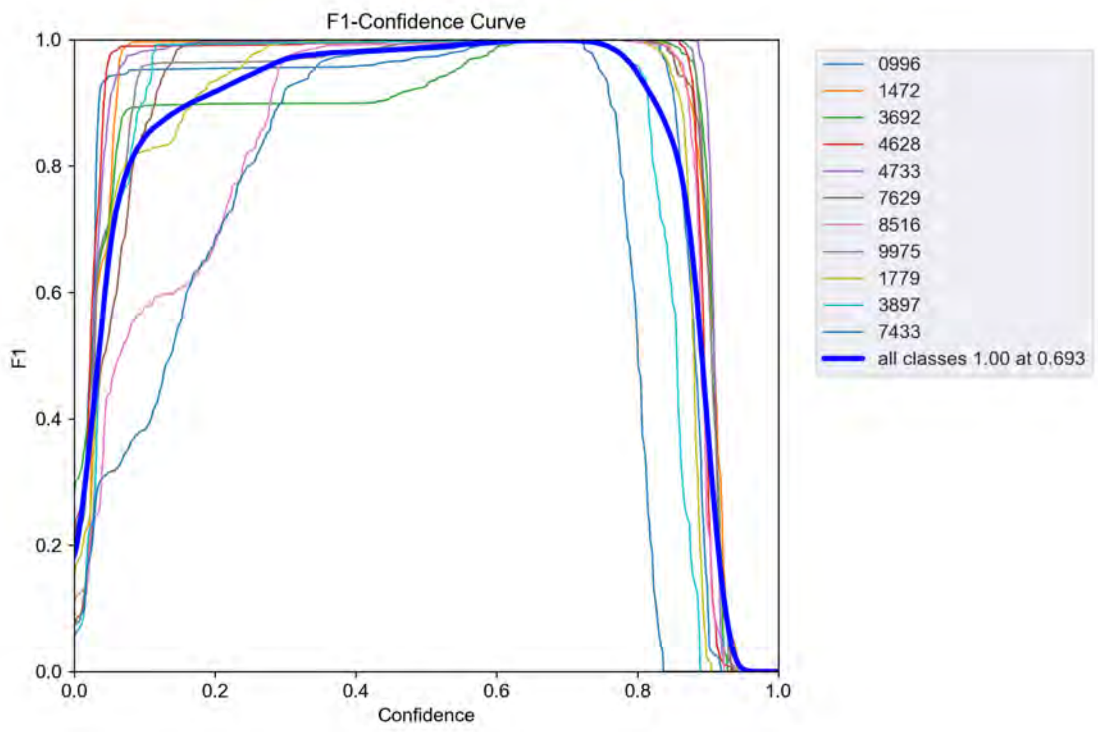


Figure A.11: 50images/class YOLOv5 Medium: F1 curve

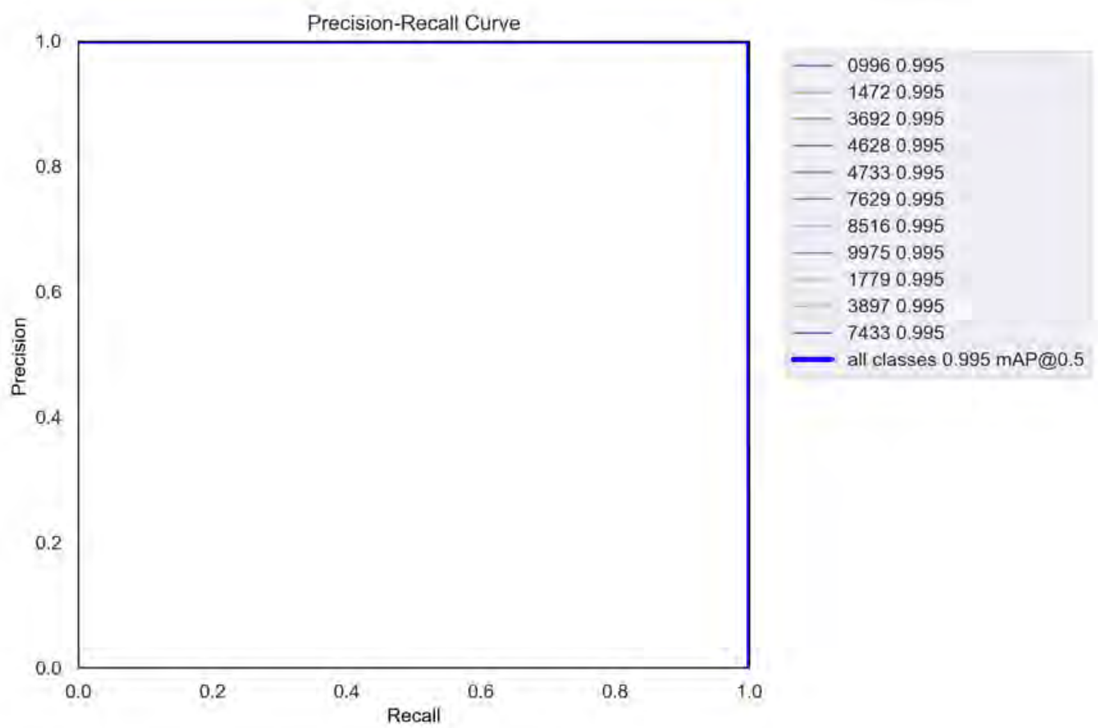


Figure A.12: 50images/class YOLOv5 Medium: precision-recall curve

A.2.2 Model trained with 70 unique images of each class

YOLOv5 Small

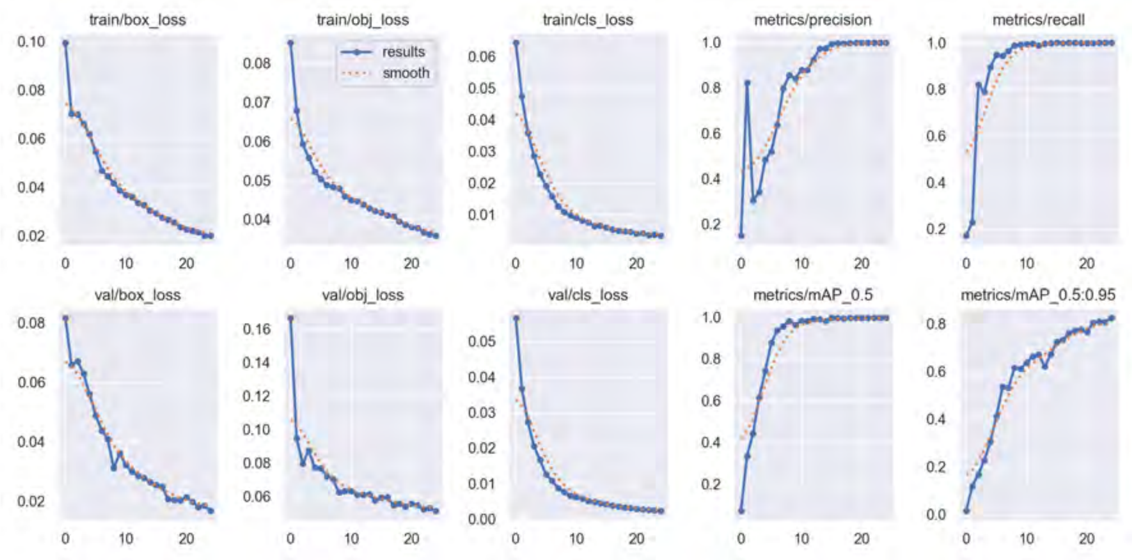


Figure A.13: 70images/class YOLOv5 Small: training metrics

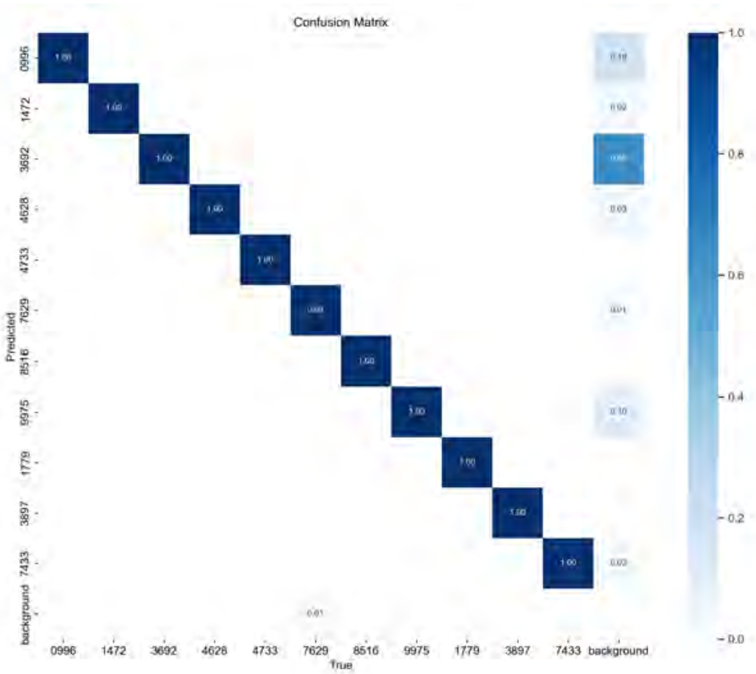


Figure A.14: 70images/class YOLOv5 Small: confusion matrix

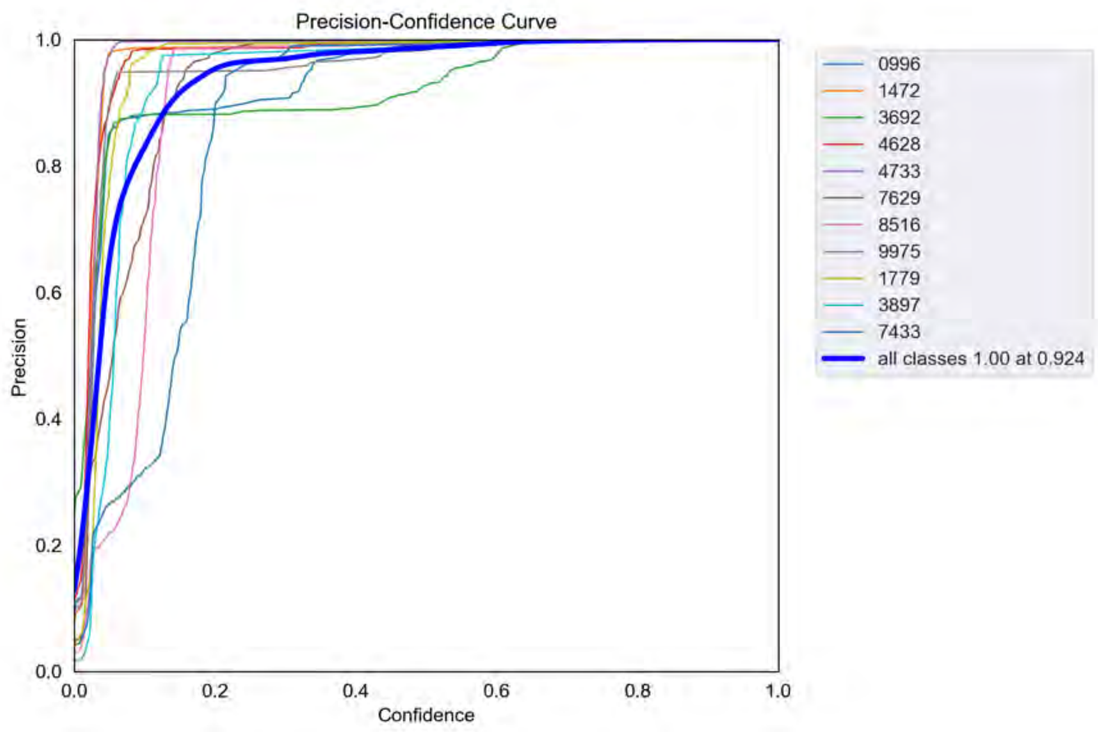


Figure A.15: 70images/class YOLOv5 Small: precision curve

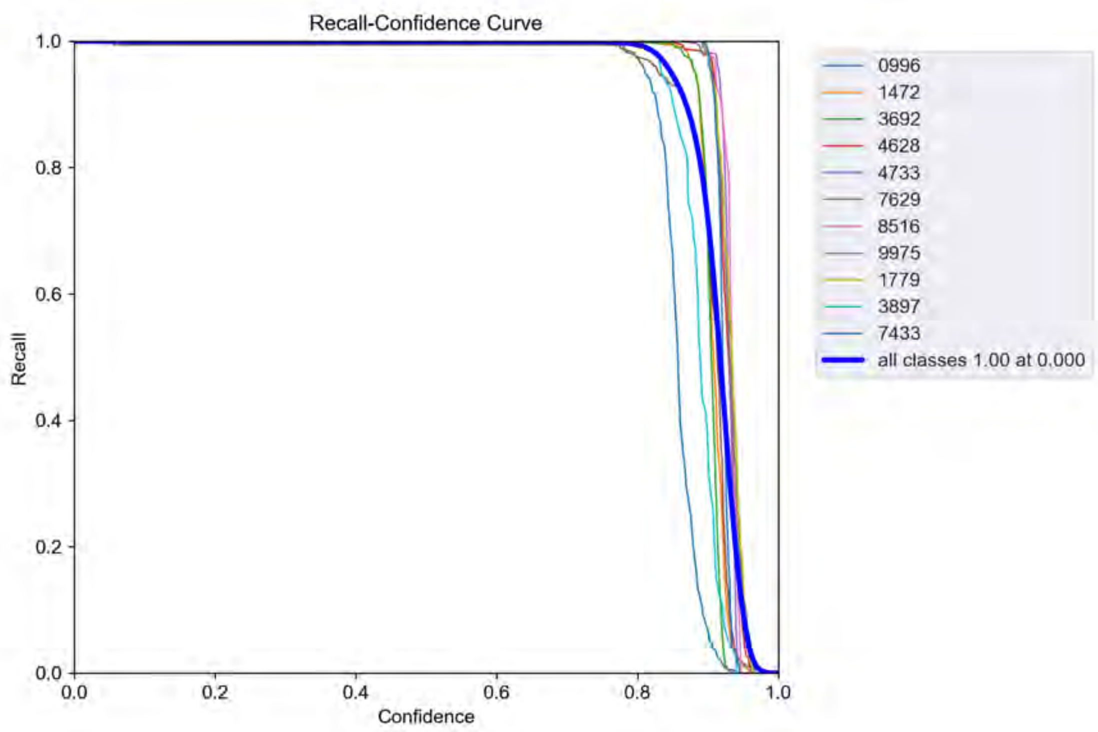


Figure A.16: 70images/class YOLOv5 Small: recall curve

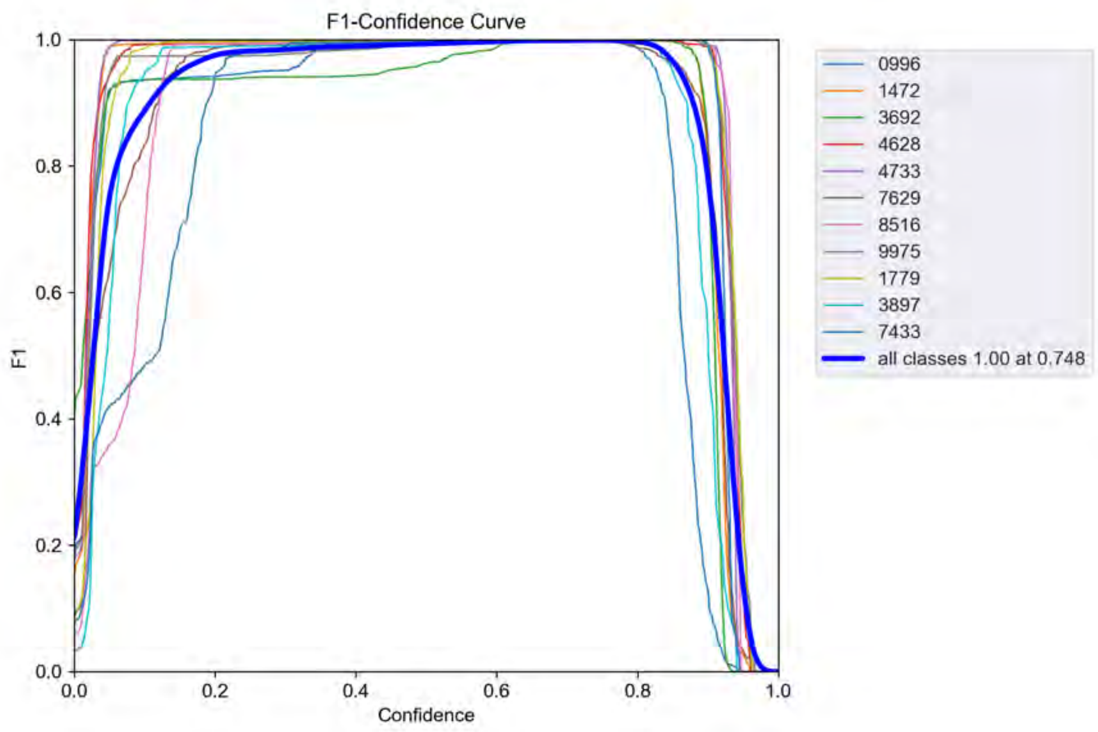


Figure A.17: 70images/class YOLOv5 Small: F1 curve

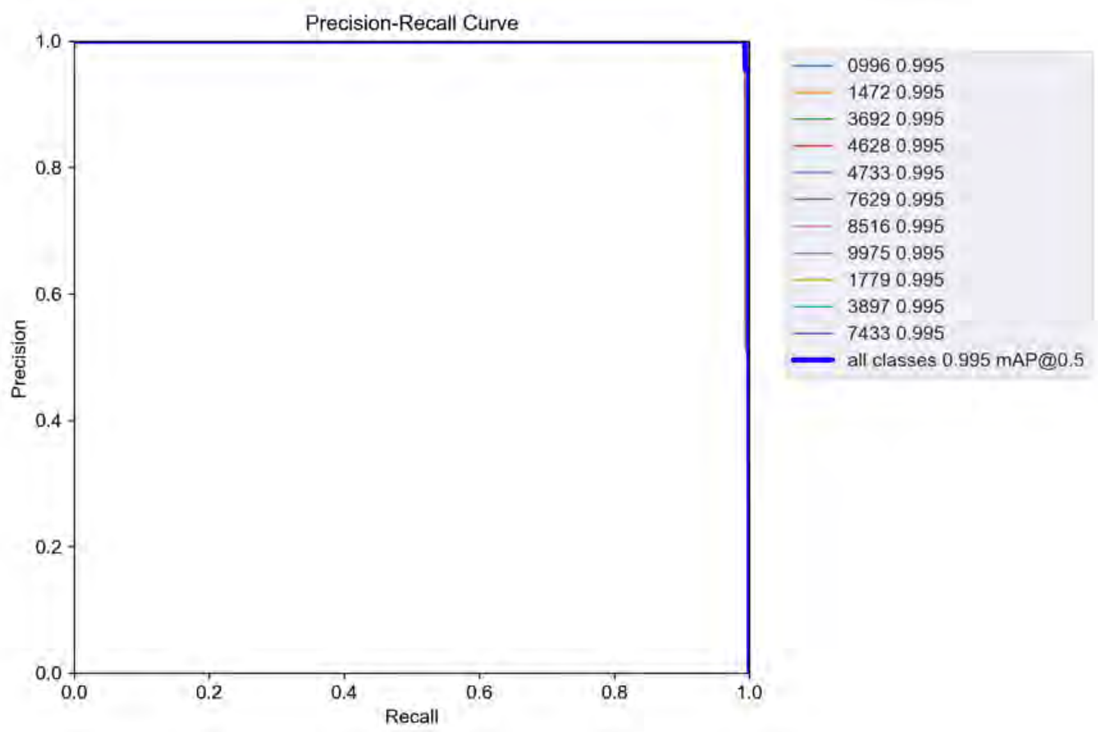


Figure A.18: 70images/class YOLOv5 Small: precision-recall curve

YOLOv5 Medium

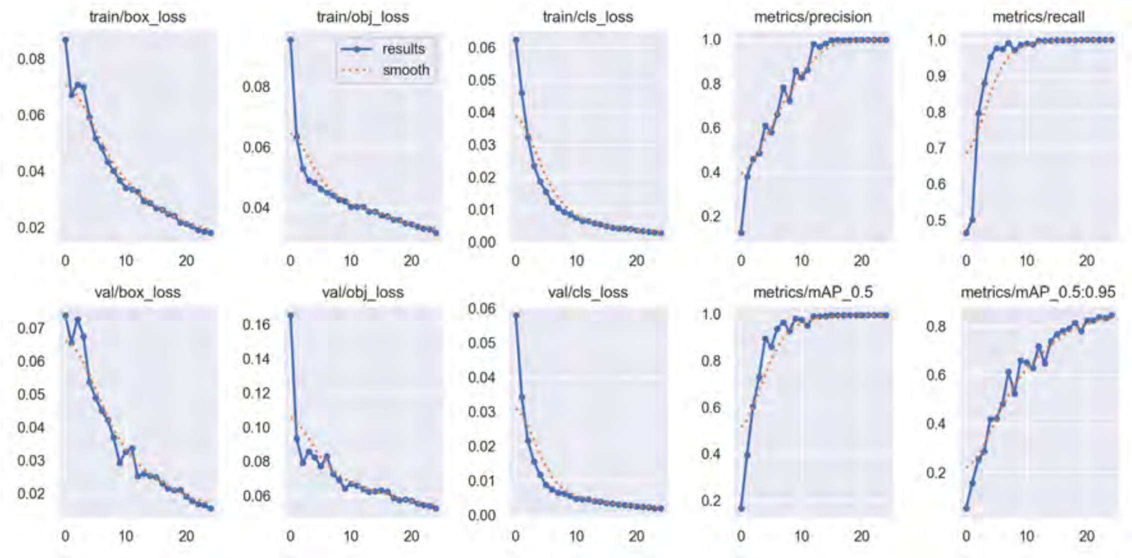


Figure A.19: 70images/class YOLOv5 Medium: training metrics

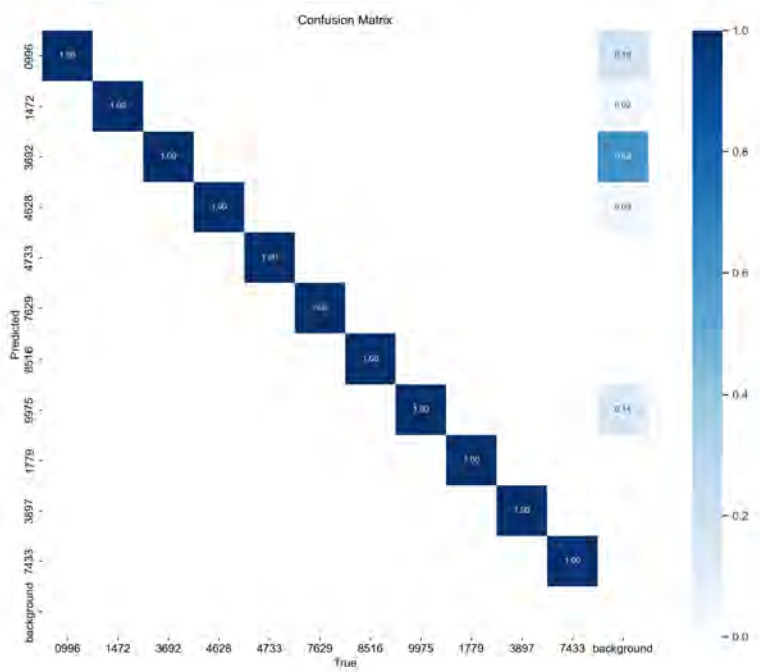


Figure A.20: 70images/class YOLOv5 Medium: confusion matrix

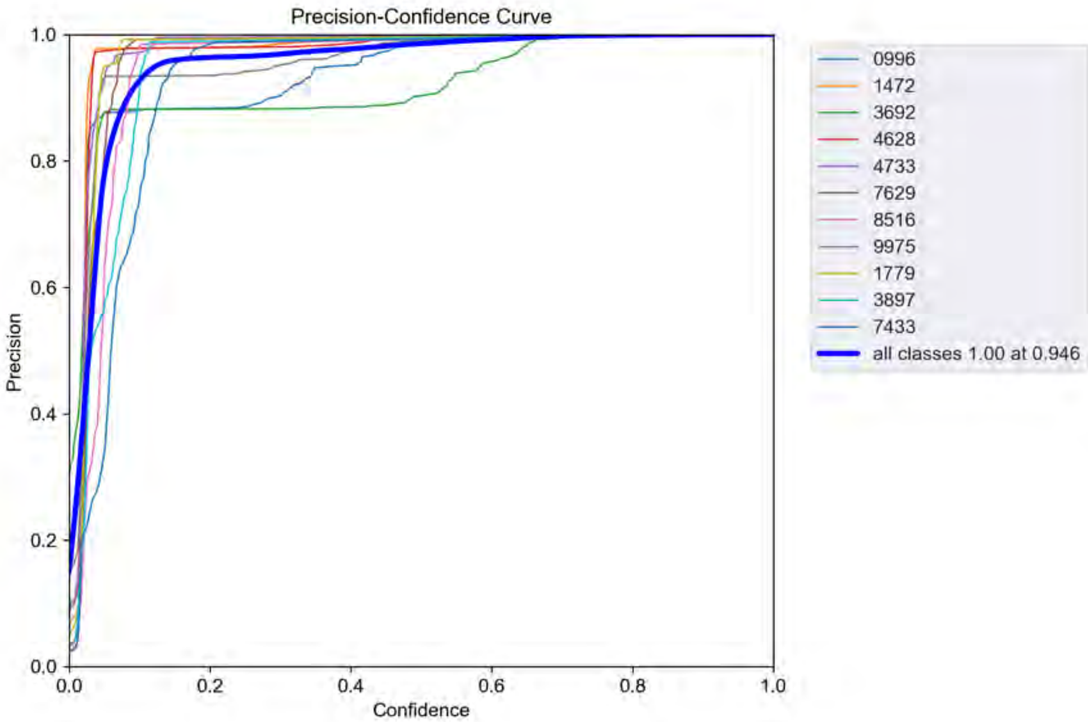


Figure A.21: 70images/class YOLOv5 Medium: precision curve

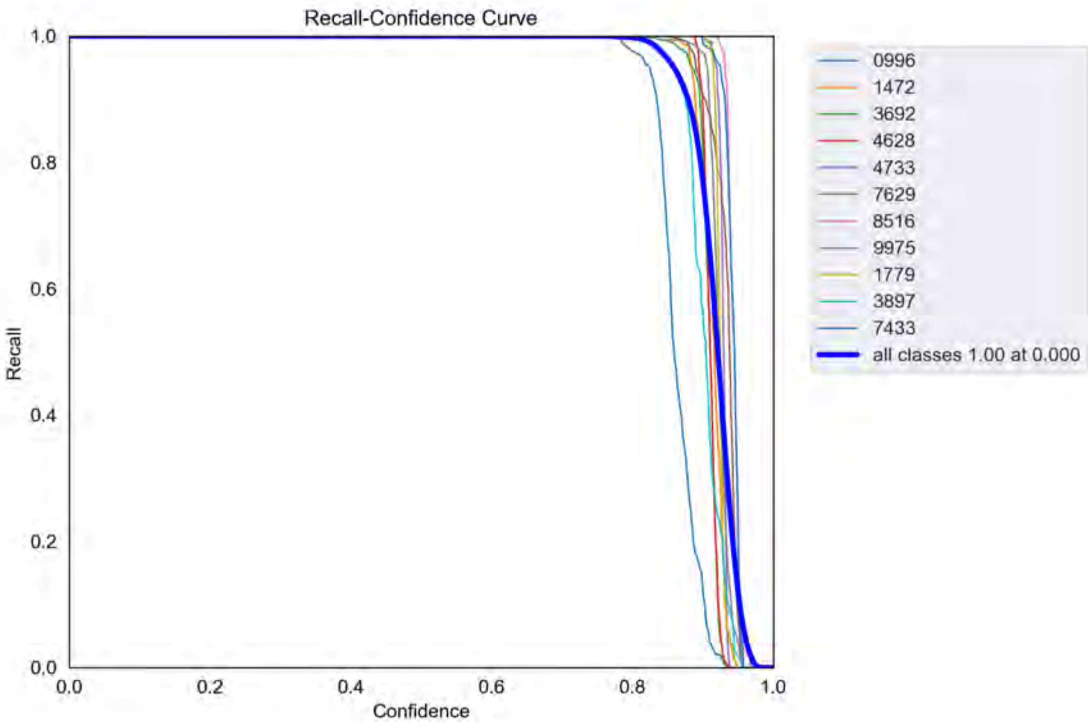


Figure A.22: 70images/class YOLOv5 Medium: recall curve

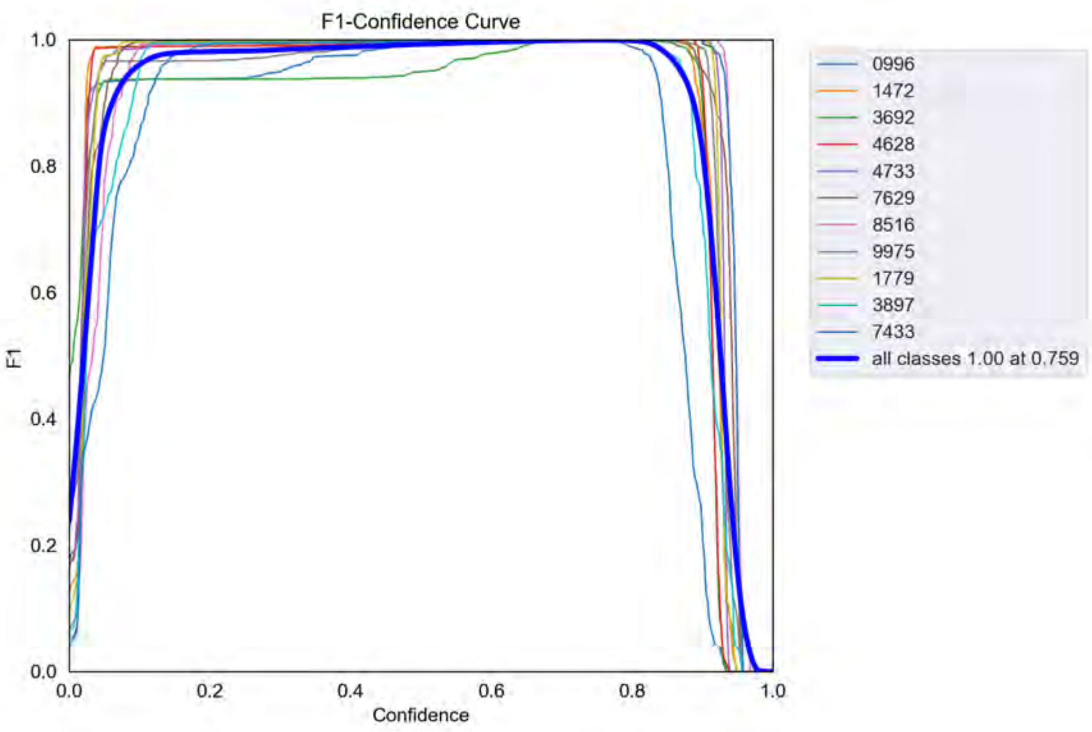


Figure A.23: 70images/class YOLOv5 Medium: F1 curve

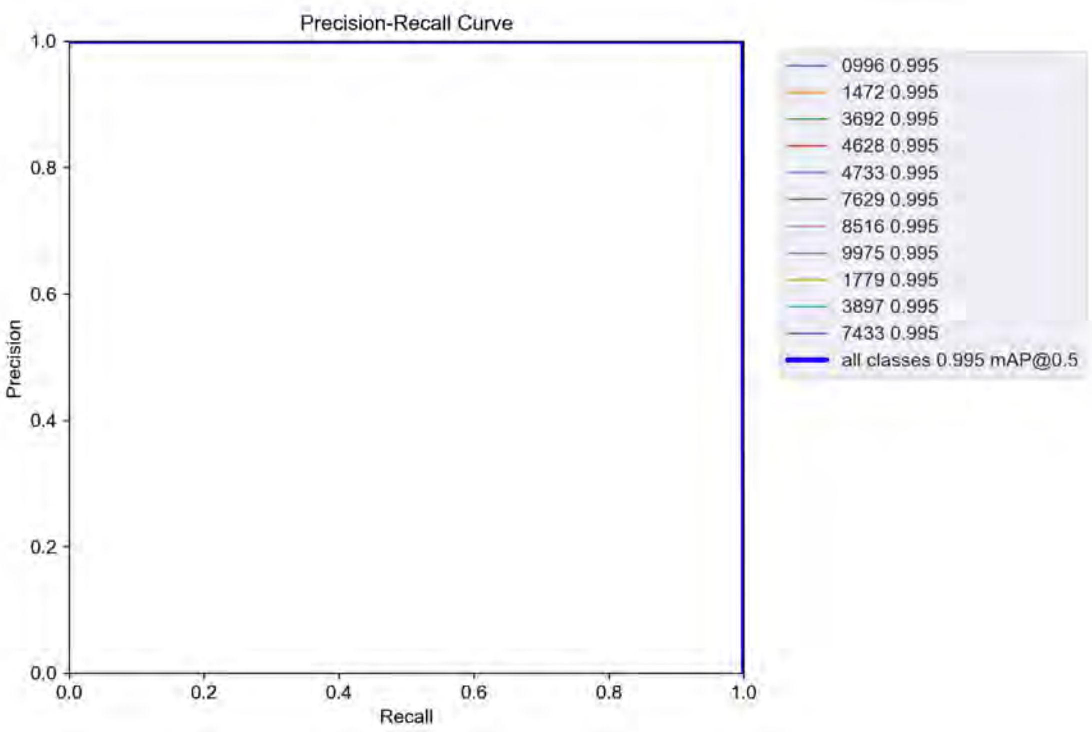


Figure A.24: 70images/class YOLOv5 Medium: precision-recall curve

A.2.3 Model trained with 90 unique images of each class

YOLOv5 Small

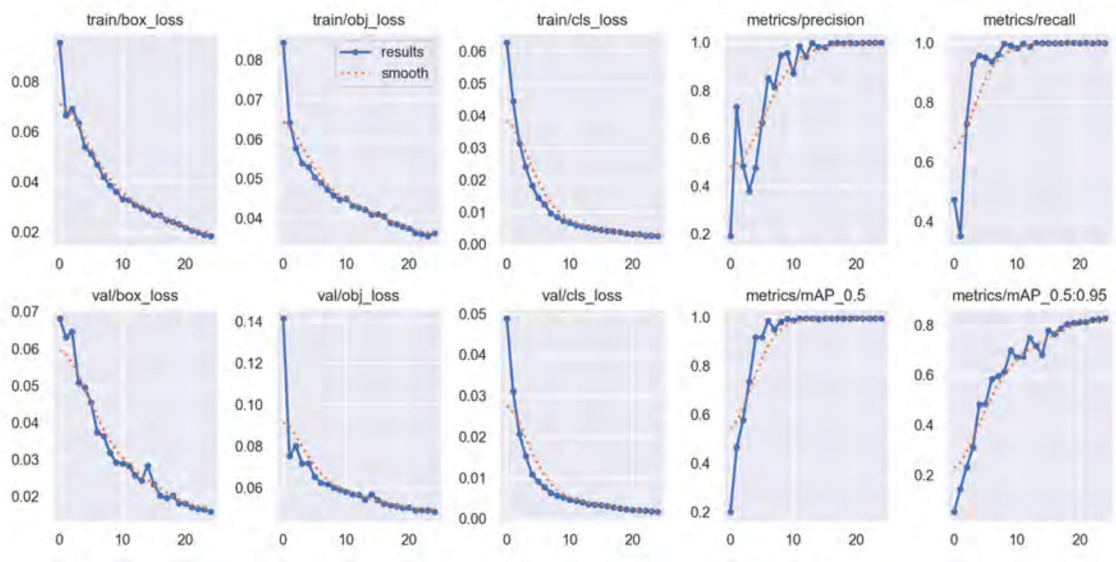


Figure A.25: 90images/class YOLOv5 Small: training metrics

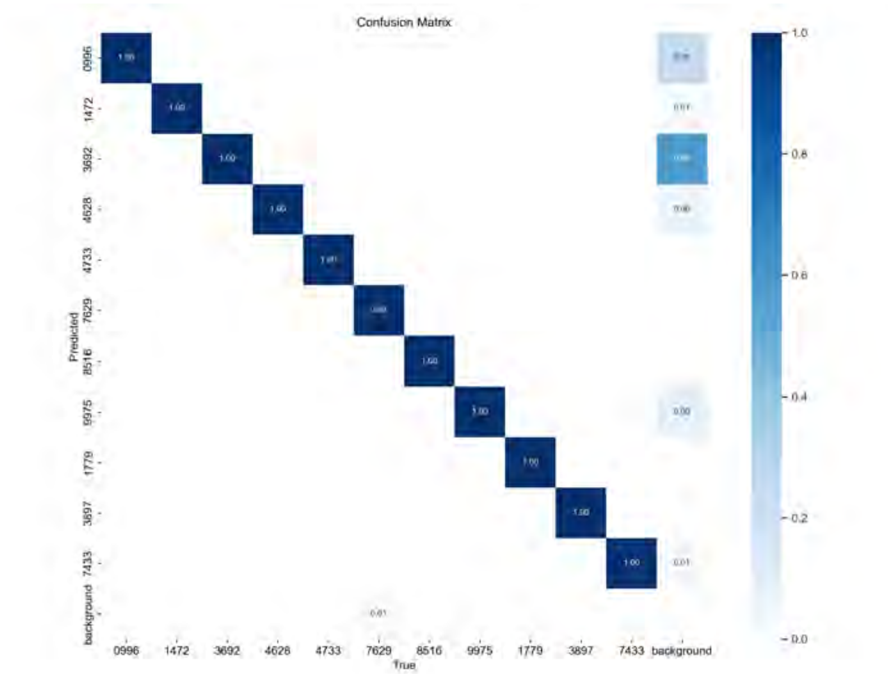


Figure A.26: 90images/class YOLOv5 Small: confusion matrix

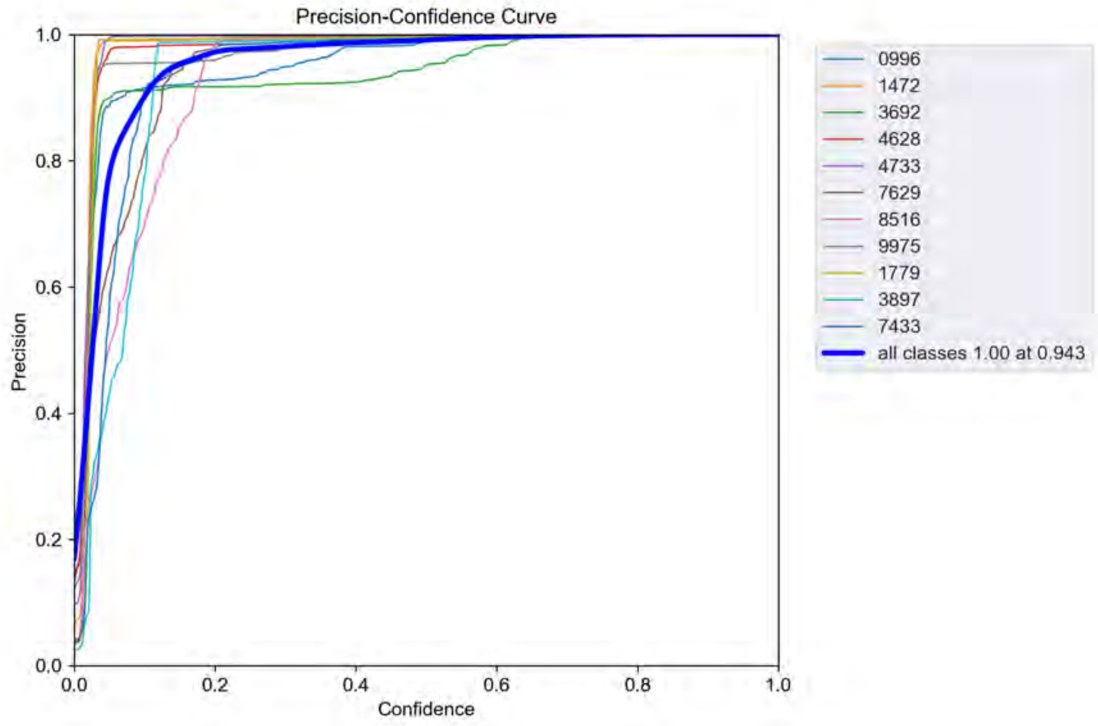


Figure A.27: 90images/class YOLOv5 Small: precision curve

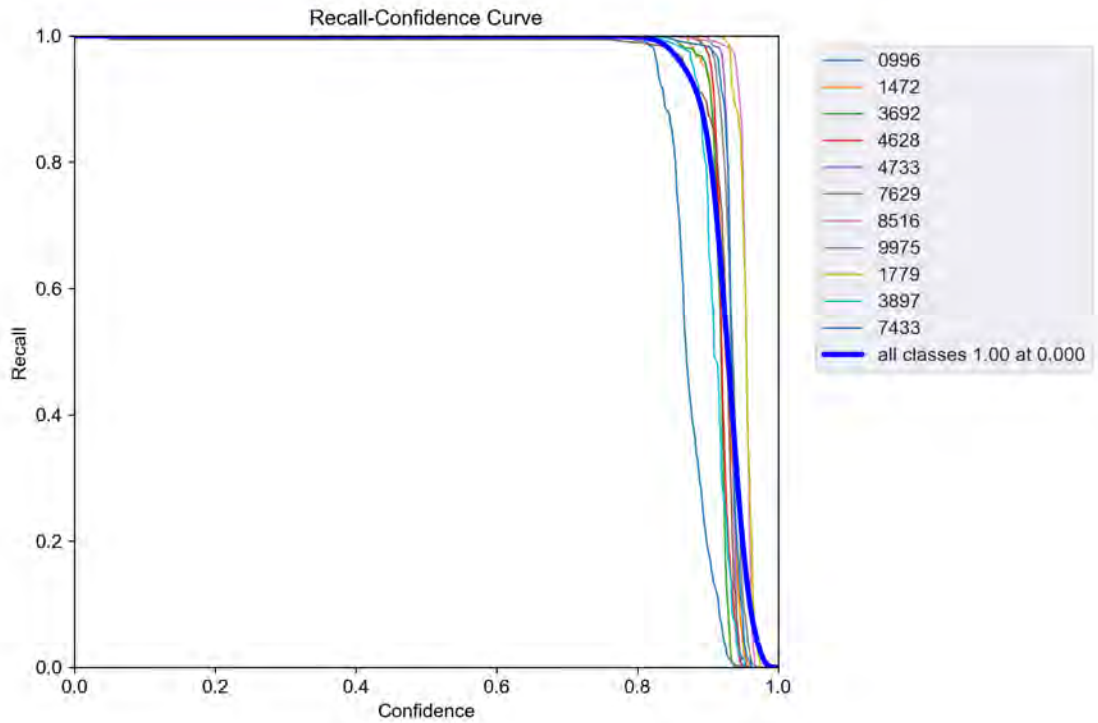


Figure A.28: 90images/class YOLOv5 Small: recall curve

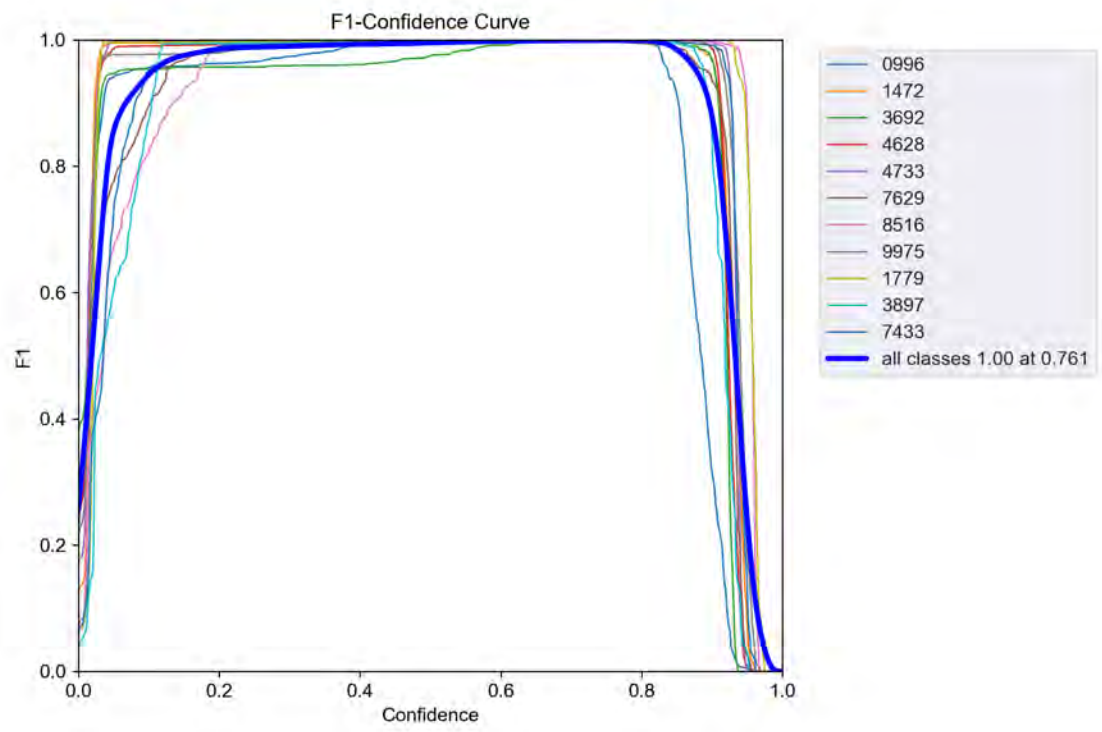


Figure A.29: 90images/class YOLOv5 Small: F1 curve

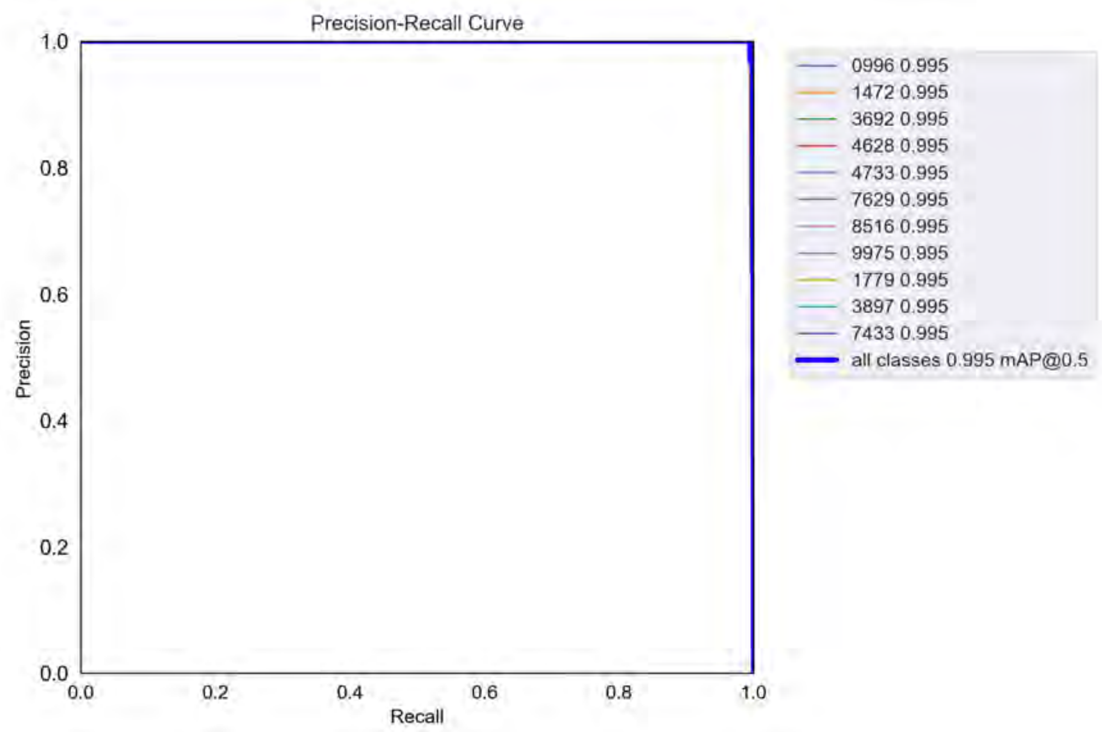


Figure A.30: 90images/class YOLOv5 Small: precision-recall curve

YOLOv5 Medium

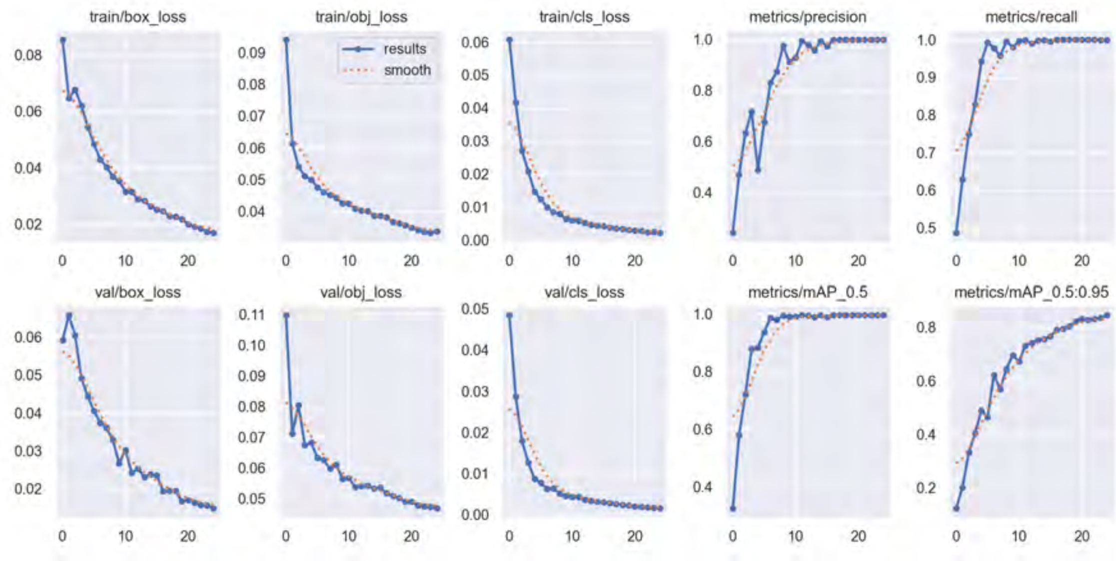


Figure A.31: 90images/class YOLOv5 Medium: training metrics

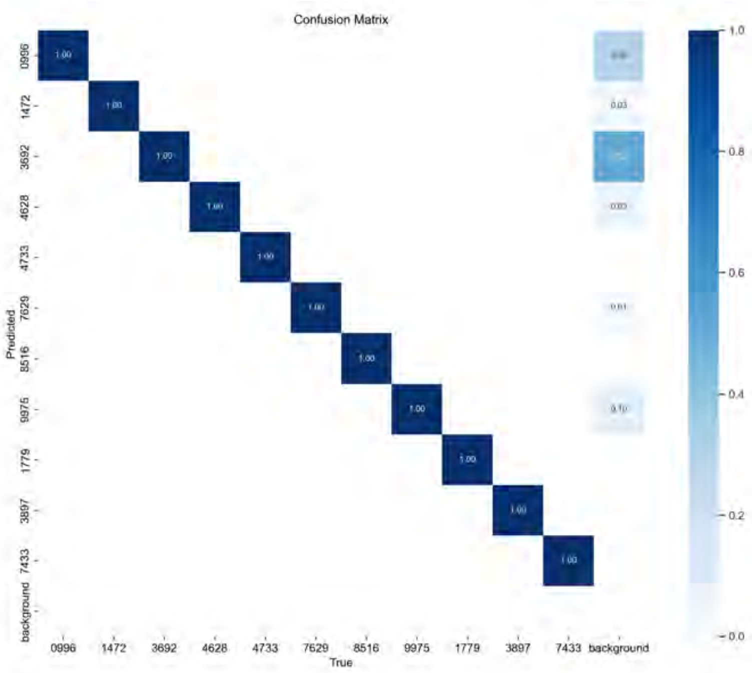


Figure A.32: 90images/class YOLOv5 Medium: confusion matrix

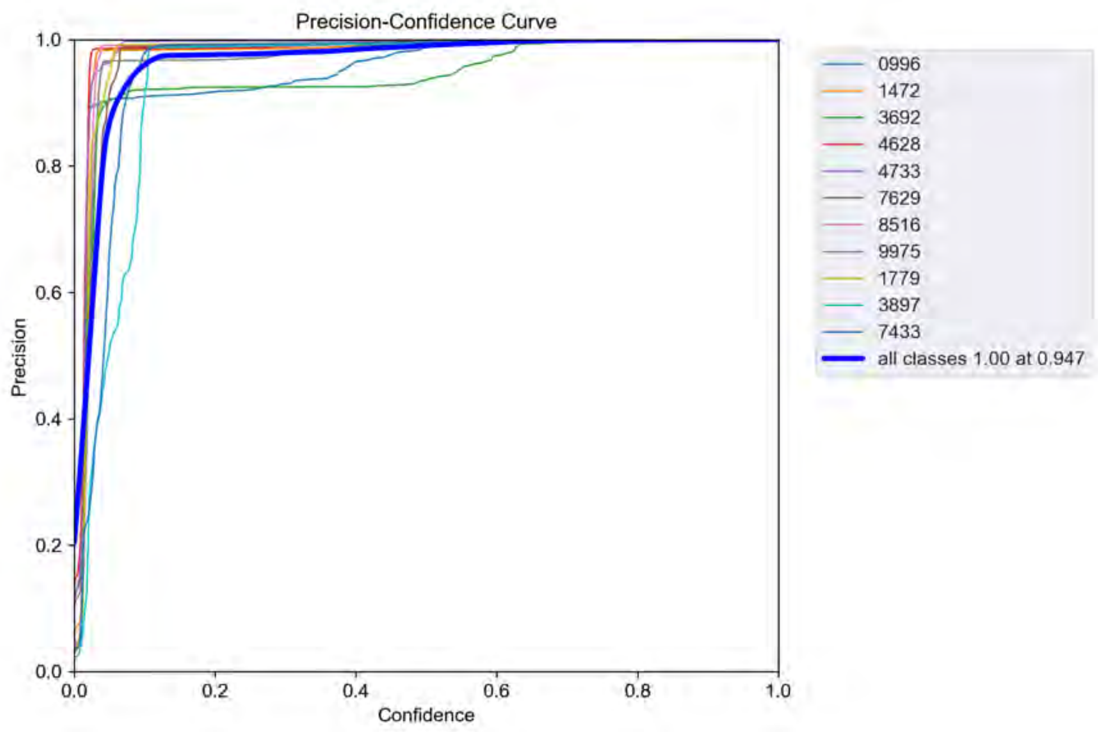


Figure A.33: 90images/class YOLOv5 Medium: precision curve

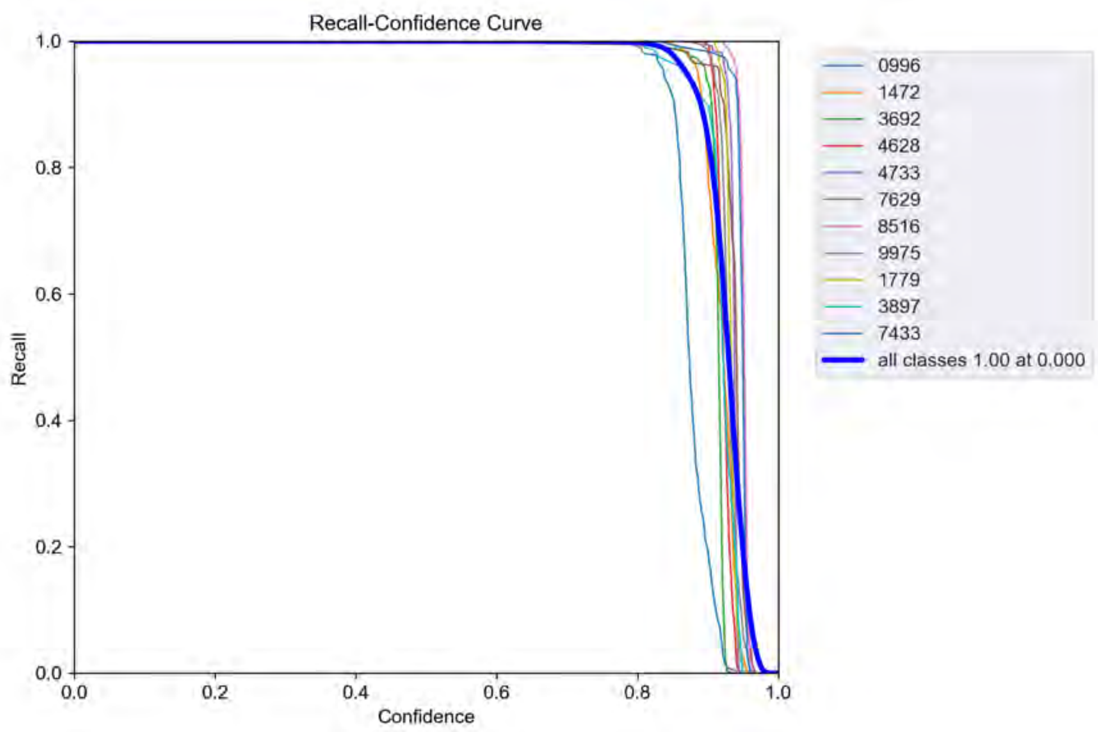


Figure A.34: 90images/class YOLOv5 Medium: recall curve

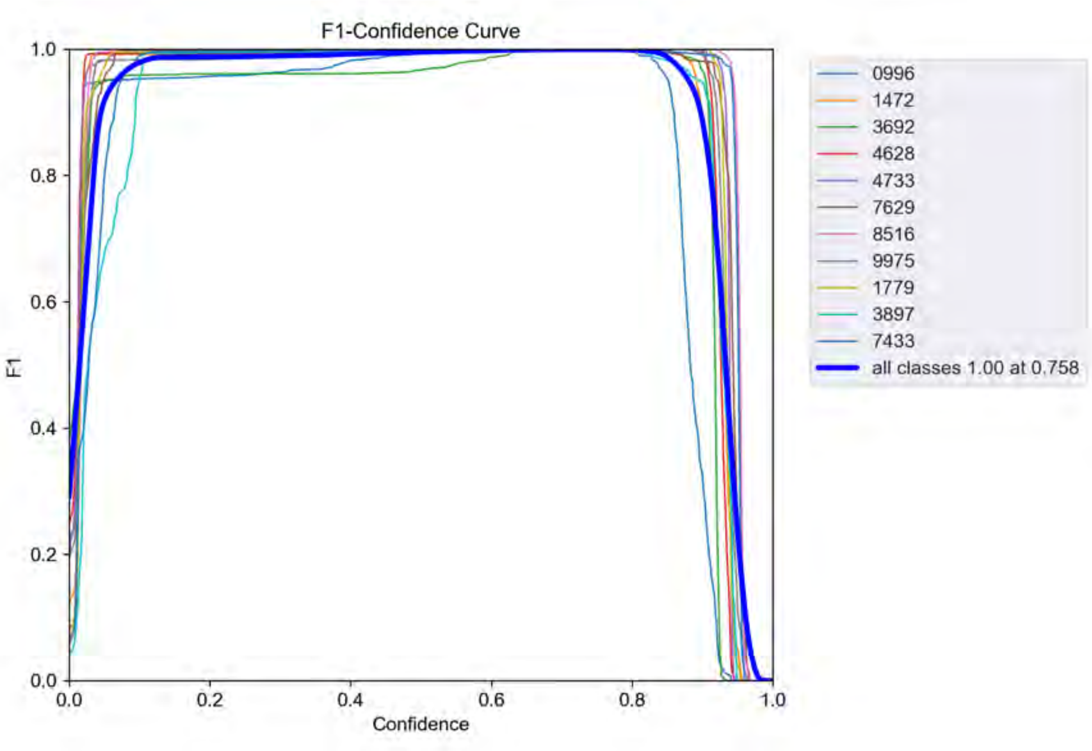


Figure A.35: 90images/class YOLOv5 Medium: F1 curve

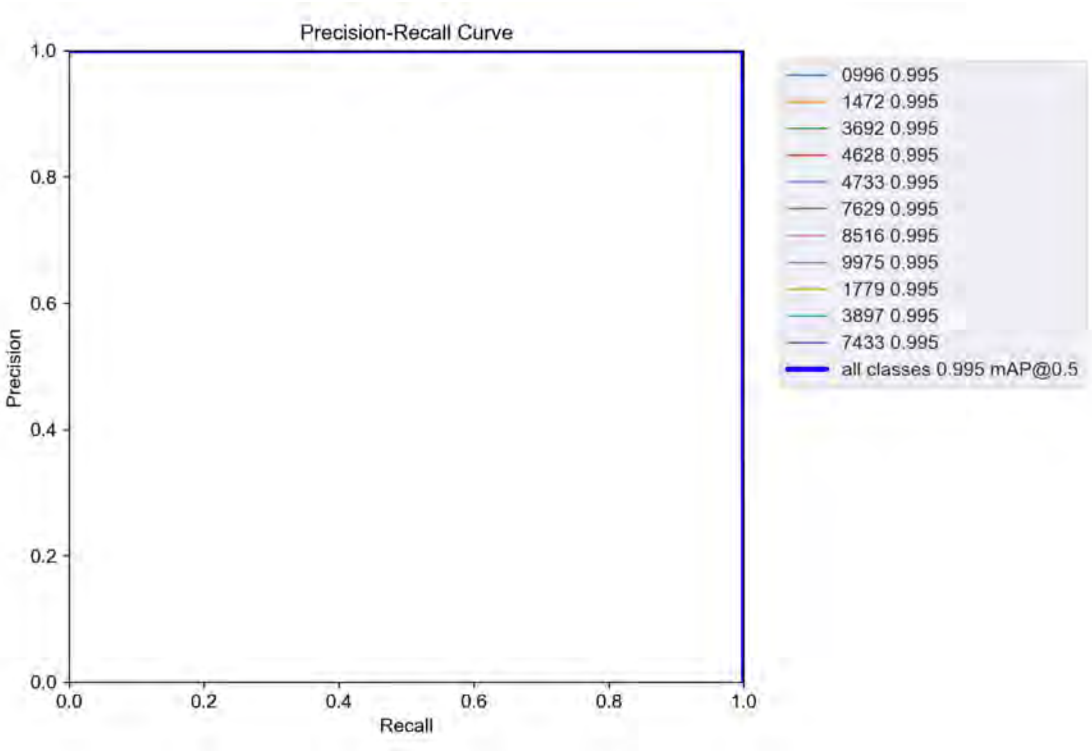


Figure A.36: 90images/class YOLOv5 Medium: precision-recall curve

DEPARTMENT OF MECHANICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY